Adventures in *Flatland*: Perceiving Social Interactions Under Physical Dynamics

Tianmin Shu[1] **(tshu@mit.edu)** **Marta Kryven**[1] **(mkryven@mit.edu)**
Tomer D. Ullman[2] **(tullman@fas.harvard.edu)** **Joshua B. Tenenbaum**[1] **(jbt@mit.edu)**
[1]Department of Brain and Cognitive Sciences, MIT
[2]Department of Psychology, Harvard University

Abstract

People make fast, spontaneous, and consistent judgements of social situations, even in complex physical contexts with multiple-body dynamics (e.g. pushing, lifting, carrying, etc.). What mental computations make such judgments possible? Do people rely on low-level perceptual cues, or on abstract concepts of agency, action, and force? We describe a new experimental paradigm, *Flatland*, for studying social inference in physical environments, using automatically generated interactive scenarios. We show that human interpretations of events in *Flatland* can be explained by a computational model that combines inverse hierarchical planning with a physical simulation engine to reason about objects and agents. This model outperforms cue-based alternatives based on hand-coded (multinomial logistic regression) and learned (LSTM) features. Our results suggest that humans could use a combination of intuitive physics and hierarchical planning to interpret complex interactive scenarios encountered in daily life.

Keywords: social perception; theory of mind; intuitive physics; Bayesian inverse planning; hierarchical planning

Introduction

We can easily read the intentions of others in their physical actions. As Oliver Wendell Holmes famously put it, *"Even a dog knows the difference between being stumbled over and being kicked."* This ease belies the understanding of physics and psychology necessary to tell the difference. More broadly, when seeing others engage in social-physical interactions (e.g. watching a soccer game) we make intuitive, fast and consistent inferences about their actions from brief observations, and without evaluative feedback. What mental mechanisms support such multi-modal and varied inference?

On one hand, the speed of social attribution suggests that it may be driven by low-level *perceptual cues*, such as facial appearance (Todorov et al., 2005; Ambady & Rosenthal, 1993), or motion (van Buren et al., 2017; Shu et al., 2018). Yet, its richness suggests a reliance on *theory of mind* (ToM), or interpreting actions of others by joint inference over incentives, abilities and goals (Gelman et al., 1995; Hamlin et al., 2013). Reasoning about physical events has likewise been studied in terms of perceptual cues (e.g.timing (Michotte, 1963) and velocity (Gilden & Proffitt, 1989)), and as driven by mentally simulating the physical world, or *intuitive physics* (Forbus, 2019; Battaglia et al., 2013).

Physical and social inferences are traditionally studied by separate empirical paradigms, since they seem to rely on different systems of knowledge (Carey, 2000), and engage different neuro-cognitive domains (Fischer et al., 2016; Sliwa & Freiwald, 2017). However, in daily life both types of attributions interact, with the interpretations of one domain relying on the understanding of the other. For example, in the classic (Heider & Simmel, 1944) experiment, the subjects' narratives illustrate both an understanding of the physical world, and of social relations and goals, such as: "The triangle is frustrated that he cannot harm the circle,". Any internal mental representation capable of accurately interpreting the nature of such multi-modal social interactions must integrate physical and social representations. This integration is necessary to differentiate between animate and physical events, and to see agents simultaneously as objects, targets of physical actions, and as agents, enacting their goals.

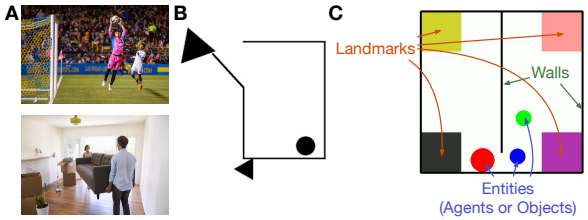

Figure 1: (**A**) Examples of real-life social interactions in physical environments: a goalkeeper blocking a shot from an opponent; two persons carrying a couch. (**B**) The classic Heider-Simmel animation abstracts such real-life interactions in animated displays of simple geometric shapes. (**C**) *Flatland* captures social scenarios, and their physical dynamics, in a controlled, procedurally generated environment. In this example subjects see three interacting circles, which represent agents and objects of different mass and size. Colored squares indicate landmarks (possible goal locations), and walls are shown by black lines. Agents and objects cannot move through walls. An agent's goal may be, for example, to move an object to a specific landmark. Agents may have relationships with other agents, expressed as goals of helping or hindering the other.

In this work we study the mechanisms of social inferences in dynamical physical scenes by introducing a new experimental paradigm, *Flatland*, inspired by Heider-Simmel animations. Several computational and quantitative studies have examined social interactions and attributions in grid-world environments (e.g. Baker et al., 2017; Kryven et al., 2016; Jara-Ettinger et al., 2015; Rabinowitz et al., 2018). *Flatland* extends these studies to a continuous physical domain, closer to the original Heider-Simmel study, but with more control over procedural stimulus generation and ground truth. Our methodology also builds on Shu et al. (2019), which used

deep reinforcement learning to generate simple social interactions in a 2D physics engine.

Flatland allows for a variety of goals, agent-to-agent relations, and physical properties of agents and objects (see Figure 1A). We interpret human attributions of goals, relations, and physical properties in this domain by a computational model that combines a hierarchical planner (based on Kaelbling & Lozano-Pérez, 2011) with a physical simulation engine[1]. Many studies explored hierarchical planning in human decision-making (e.g. Balaguer et al., 2016; Huys et al., 2015), and recently in social inference (Yildirim et al., 2019). We show that a combined hierarchical planning and physical engine model outperforms cue-based alternatives, such as multinomial logistic regression and LSTM, in predicting human interpretations of ambiguous events. Our results suggest the role of complex abstract physical and mentalistic concepts in social inference.

Computational Modeling

Flatland

Flatland is a 2D simulated physical environment with multiple interacting agents and objects. Agents can have two types of goals: (1) a personal goal $g_i \in \mathcal{G}$, and (2) a social goal of helping or hindering another agent. Agents are subject to physical forces, and can exert self-propelled forces to move. Agents can also attach objects to their bodies and release them later. In the current study, agents have accurate and explicit knowledge of the other agents' goals. However, the *Flatland* environment can be extended to scenarios with incomplete information.

Formally, agents are represented by a decentralized Multi-agent Markov Decision Process (Dec-MDP), i.e., $\langle \mathcal{S}, \mathcal{A}, R_i, \mathcal{T}_i \rangle$, $\forall i \in N$, where N is the number of agents, $\mathcal{S}$ and $\mathcal{A}$ are the state set and the action set shared by all agents, $R_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the agent's reward function, $\mathcal{T}_i : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the agents' state transition probability. The amount of force an agent can exert, f_i, defines the agent's strength, and shapes its dynamics in the physical environment. An agent's state transition probability $\mathcal{T}_i$ can be written as $P(s'|s, a, f_i, \theta_i)$, where θ_i denotes the physical properties of the agent, other than its strength, (e.g. mass and shape). Assuming that all bodies have the same density, θ_i is easily observable based on visual appearance, but this assumption can be relaxed.

An agent's reward is jointly determined by: (1) the reward of its own goal, (2) the reward of other agents' goals, (3) its relationships with other agents, and (4) the cost of actions. Formally, we define an agent's reward as:

$$R_i(s,a) = R(s,g_i) + \sum_{j \neq i} \alpha_{ij} R(s,g_j) - C(a), \qquad (1)$$

where $C(a)$ is the cost function; α_{ij} indicates agent i's relationship with agent j, including how much agent i cares about agent j's goal. For a friendly relationship, $\alpha_{ij} > 0$; for an adversarial one, $\alpha_{ij} < 0$; and $\alpha_{ij} = 0$ if the relationship is neutral.

Given this physical and social setup, we now consider how an agent could plan to achieve its goals in this environment. Interpreting an agent's actions would then require the inversion of this planning process. A classic MDP-based approach would prove exceedingly costly in the continuous physics of *Flatland*, coupled with the agents' composite rewards. In this work we deal with this complexity by incorporating a hierarchical planner inspired by the task and motion planning (TAMP) framework (Kaelbling & Lozano-Pérez, 2011).

Hierarchical Planning

Figure 2 shows how the hierarchical planner (HP) works. Given an agent i's goal, strength, relationships with other agents, and the other agents' goals, HP generates the best action to take at any given state. An agent can pursue its own goal, or the goal of another agent. HP searches for plans $\Pi_{ij} = \{a^\tau\}_{\tau=t}^{t+T-1}$ with a finite horizon of T steps for all goals g_j, $\forall j \in N$. Any g_j such that $j \neq i$ is a goal of another agent j. Each plan is simulated using the physics engine, and the agent's cumulative reward following that plan is given by a composite value function, $V(\Pi_{ij}) = \sum_{\tau=0}^{T-1} R_i(s^{t+\tau}, a^{t+\tau})$, which incorporates the reward of its personal goal and the weighted rewards of other agents' personal goals.

The plan with the highest cumulative reward is selected as the final plan generated by the HP. To better adapt to other agents' plans, in the current implementation HP returns only the first action of the selected plan, and re-plans by searching for new plans at every step. So, the plans can be frequently adjusted according to the latest state.

To generate an optimal plan for each possible goal, the HP adopts a two-level architecture. First, a Symbolic Planner (SP) prepares a sequence of sub-goals for a given goal. This entails generating symbolic states from physical states[2], and creating a sequence of sub-goals that reaches the final goal. For example, a sub-goal could entail grabbing an object, blocking a door, or moving to a specific location. In the present study, SP used A^* search to find the shortest path to the goal in the space of symbolic states. Second, a Motion Planner (MP) generates a sequence of actions[3] that achieves each sub-goals using Monte-Carlo Tree Search (MCTS) together with a physics engine. Note that alternative implementations of SP and MP may be suitable in different domains.

Formally, let $\pi^o(o_i^t | s^t, f_i, \{g_j\}_{j\in N}, \{\alpha_{ij}\}_{j\neq i}) \propto e^{V(\Pi_{ij})}$ be the plan selection policy, where $o_i^t \in \{g_k\}_{k\in N}$ is the selected goal, and let $\pi(a_i^t | s^t, f_i, o_i^t)$ be the policy for the selected goal, computed by MCTS. Then, the agent's final policy is:

$$a_i^t \sim \pi_(a_i^t | s^t, f_i, \{g_j\}_{j\in N}, \{\alpha_{ij}\}_{j\neq i}), \qquad (2)$$

[1] https://github.com/pybox2d/pybox2d

[2] The symbolic states are predefined predicates: *On(object, landmark)*, *Reachable(agent, object)*, *Attached(agent, object)* and their negation).

[3] Here the actions are forces that can be applied in eight possible directions, grabbing or releasing an object, stopping, and no force.

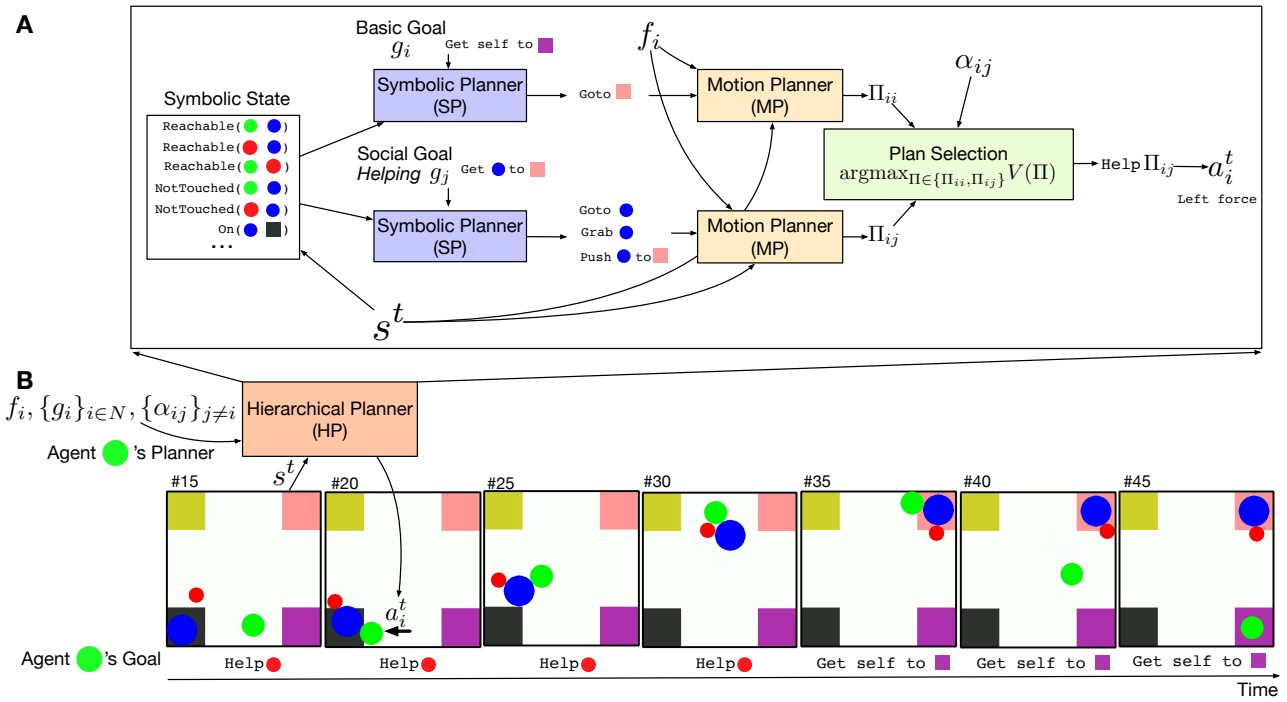

Figure 2: An illustration of an agent's Hierarchical Planner (HP).(**A**) Using Symbolic Planner (SP) and Motion Planner (MP), HP optimizes a value function that combines a personal and a social goal (helping or hindering another agent). In this example the agent has a helping social goal. (**B**) Each agent replans its actions every a few steps in response to a changing state of other agents and objects. An agent may switch goals, when the expected reward of the new goal outweighs that of the current goal. For example, the green agent first helps the red agent to bring the blue object to the pink landmark; once it has been reached, the green agent switches to pursuing its personal goal.

where

$$\pi(a_i^t|s^t, f_i, \{g_j\}_{j\in N}, \{\alpha_{ij}\}_{j\neq i}) = \sum_{o_i^t\in\{g_k\}_{k\in N}} \pi^a(a_i^t|s^t, f_i, o_i^t)\pi^o(o_i^t|s^t, f_i, \{g_j\}_{j\in N}, \{\alpha_{ij}\}_{j\neq i}). \tag{3}$$

Attributing Goals, Relationships, and Strengths

By running the HP forward we can generate arbitrary interactive scenarios in the *Flatland* environment. Such scenarios could involve a number of objects of different sizes, shapes, and appearances, and a number of agents with personal and social goals. People viewing animations of this kind tend to interpret them in terms of a narrative about the agents' relationships, incentives, and abilities (Heider & Simmel, 1944). Such interpretations may arise either from identifying specific cues, or from applying a more structured, theory-like understanding of objects and agents. We formalize these two views of human judgement using cue-based models as well as a theory-based generative model that relies on Bayesian inverse planning enabled by the HP and physics simulation.

Generative Social and Physical Inference (GSPI) GSPI conducts Bayesian inference of latent variables (i.e., agents' goals, relationships, strengths) to describe an observed social interaction through a generative model consisting of our heirarchical planner and a physics engine. For each hypothesis of the latent variables, GSPI i) samples optimal plans w.r.t. Eq. 2, and ii) simulates entities' trajectories in the physics engine based on the hypothesis as well as the sampled plans. GSPI then defines the likelihood of the hypothesis by how much the simulated trajectories deviate from the observed trajectories. Combined with the priors of the latent variables, GSPI computes the posterior of the hypothesis using Bayes's rule:

$$P(g_i, g_j, f_i, f_j, \alpha_{ij}, \alpha_{ji}|s_i^{1:T}, s_j^{1:T}) \propto P(s_i^{1:T}, s_j^{1:T}|g_i, g_j, f_i, f_j, \alpha_{ij}, \alpha_{ji}) \cdot P(g_i, g_j, f_i, f_j, \alpha_{ij}, \alpha_{ji}). \tag{4}$$

Given this principle, we show how GSPI can be used for inferring agents' goal selection, relationships, and comparing their relative strengths in details as fellows.

First we can calculate the posterior probability of the agents' goals, given observations and given the agents' social and physical properties:

$$P_{ij}(o_i^t, o_j^t) = P(o_i^t, o_j^t|s^t, s^{t+1}, g_i, g_j, f_i, f_j, \alpha_{ij}, \alpha_{ji}) \propto \sum_{a_i^t, a_j^t} P(s^{t+1}|s^t, a_i^t, a_j^t)\pi^a(a_i^t|s^t, f_i, o_i^t)\pi^o(o_i^t|s^t, f_i, \{g_i, g_j\}, \alpha_{ij}) \cdot \pi^a(a_j^t|s^t, f_j, o_j^t)\pi^o(o_j^t|s^t, f_j, \{g_i, g_j\}, \alpha_{ji})P(a_i^t)P(a_j^t)P(o_i^t)P(o_j^t), \tag{5}$$

where $P(s^{t+1}|s^t, a_i^t, a_j^t) = e^{-\beta||s^{t+1}-\hat{s}^{t+1}||}$, with $\hat{s}^{t+1}$ being the predicted next state after taking a_i^t based on physics simulation. Here β controls the agent's proximity to the optimal plans generated by the HP. A large β means that the agent will follow the optimal plan; as β becomes smaller, it is increasingly likely to deviate from the optimal plan. $P(o)$ and $P(a)$

are uniform priors.

The probability that both agents are pursuing their personal goals in the last T steps is given by:[4]

$$P(o_i = g_i, o_j = g_j | s^{1:T}) = \sum_{f_i, f_j, \alpha_{ij}, \alpha_{ji}} \prod_t P_{ij}(g_i, g_j) P(f_i) P(f_j) P(\alpha_{ij}) P(\alpha_{ji}). \quad (6)$$

The probability that agent i is pursing a social goal (helping or hindering another agent) is given by:

$$P(o_i = g_j, o_j = g_j | s^{1:T}) = \sum_{g_i, f_i, f_j, \alpha_{ij}, \alpha_{ji}} \prod_t P_{ij}(g_j, g_j) P(g_i) P(f_i) P(f_j) P(\alpha_i) P(\alpha_j). \quad (7)$$

Here we discretize f and α for making the computations tractable. We assume uniform priors for goals and strengths. For α, we assume that $P(\alpha = 0) = P(\alpha > 0) = P(\alpha < 0) = 1/3$, so that there is no bias toward any type of relationship.

To infer the relationship between two agents, we first derive the posterior probabilities of α_{ij} and α_{ji}.

$$P(\alpha_{ij}, \alpha_{ji} | s^{1:T}) = \prod_t \sum_{g_i, g_j, f_i, f_j} \sum_{o_i^t, o_j^t} P_{ij}(o_i^t, o_j^t) P(g_i) P(g_j) P(f_i) P(f_j) P(\alpha_{ij}) P(\alpha_{ji}). \quad (8)$$

Based on Eq. 8, we can derive the posterior probability of specific relationships. E.g.,

$$P(\text{Adversarial} | s^{1:T}) = \sum_{\alpha_{ij} \le 0, \alpha_{ji} < 0} \sum_{\alpha_{ij} < 0, \alpha_{ji} \le 0} P(\alpha_{ij}, \alpha_{ji} | s^{1:T}). \quad (9)$$

Finally, the expected strength difference between two agents is given by:

$$\mathrm{E}[f_i - f_j | s^{1:T}] = \sum_{f_i} \sum_{f_j} (f_i - f_j) P(f_i, f_j | s^{1:T}), \quad (10)$$

where

$$P(f_i, f_j | s^{1:T}) = \prod_t \sum_{g_i, g_j, \alpha_{ij}, \alpha_{ji}} \sum_{o_i^t, o_j^t} P_{ij}(o_i^t, o_j^t) P(g_i) P(g_j) P(f_i) P(f_j) P(\alpha_{ij}) P(\alpha_{ji}). \quad (11)$$

Cue-based models We compared the GSPI model with three cue-based alternatives: *Cue-based-1*: Multinomial logistic regression based on feature statistics of the whole video; *Cue-based-2*: Multinomial logistic regression based on concatenated feature statistics of chunks of the video; *Cue-based-3*: Long short-term memory (LSTM). Each cue-based model was trained on 400 stimuli, not used in the experiment. Following (Ullman et al., 2010), we used the following cues for each agent: (1) coordinates, (2) velocity, (3) acceleration, (4) relative velocity w.r.t. other entities and landmarks, (5) distance to other entities and landmarks, (6) whether the agent is touching another entity. The LSTM model accepts the sequence of these cues as input and learns motion features by itself. For logistic regression models, we encode the cue sequences as statistics (mean, minimum, maximum, standard deviation) to obtain motion features. *Cue-based-1* used the statistics over the whole video as input, and *Cue-based-2* concatenated statistics of short chunks of a video as input. To train these models, we generated 400 training videos by randomly sampling agents' goals, relations, and strengths as well as the environment layout, sizes and initial positions of entities. Note that these 400 training videos were not shown in the human experiment.

[4]Note that the equations here are constrained to two-agent scenarios for simplicity and readability, but they can be easily extended to more general cases.

Methods

Flatland is a simple but rich environment, capable of generating many visually distinct scenes from a relatively small number of underlying physical and social variables. *Flatland* scenarios allow us to quantitatively test alternative accounts of human physical and social reasoning. We have described two such basic alternatives – i) a theory-like inference that requires forward planning models and physical simulation for the physical and psychology of agents in *Flatland*, and ii) a cue-based alternative that relies on many separate visual cues to map between observed social interactions and agents' goals, relationships and strengths. We next describe an empirical study of human inferences in *Flatland*, in order to assess the fit of these two different models.

Procedure

The experiment[5] was presented in a web browser using psiTurk (Gureckis et al., 2016). The instructions explained how *Flatland* works. After reading instructions, subjects completed 3 comprehension quizzes for judging goals, relationship, and strengths respectively. Subjects who failed to accurately respond to all quizzes were asked to read the instructions again until they correctly responded to all quizzes. Next, subjects responded to two practice stimuli, similar to the stimuli presented in the main experiment, the responses to which were not included in the analysis. After completing the practice, subjects saw 6 stimuli and reported: (1) the goals of each agent, (2) the relationship between agents, and (3) the relative strength of each agent. The responses were given by selecting the appropriate items from a multiple-choice list.

Subjects

120 subjects (mean age = 38.4; 45 female) were recruited on Amazon Mechanical Turk and paid $1.60 for 12 minutes. Subjects gave informed consent. The study was approved by the MIT Institutional Review Board.

[5]The exact experimental setup (screenshots) can be found at https://osf.io/25nsr/?view_only=ce34eb376d0c4f3dbf3a095bd7dafb60

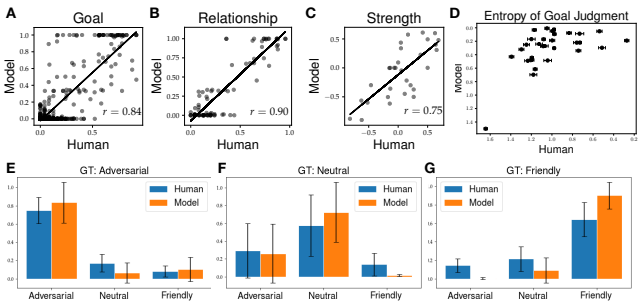

Figure 3: Comparing the GSPI model's inferences to human responses. (**A**) Probabilities of each of the possible agent goals given by the model, plotted against the averaged human responses. (**B**) Probabilities of each type of relationship (Neutral, Friendly, Adversarial) given by the model plotted against the averaged human responses. (**C**) The model's estimate of the strength difference between the two agents against the averaged human response. (**D**) Entropy (in bits) of human goal judgements plotted against the entropy of the model's goal judgments. Each data point represents one stimulus, and the error bars indicate the bootstrapped 95% confidence intervals. Notably, stimuli that exhibited higher entropy (more ambiguity) in humans were also harder for the model. (**E**-**G**) Human and model's inferred agents' relationships, given ground-truth. Error bars show 95% confidence intervals. Both humans and model correctly identified the relationships in most of the stimuli, and had a higher degree of confusion when the ground-truth relationship was neutral.

Stimuli

Stimuli were 30 *Flatland* animations[6], which always contained three interacting bodies (shown as circles) and four landmarks (squares placed at the four corners of the screen, as shown in Figure 1C). Two of the interacting bodies were always agents and one was an object. Subjects were informed which of the circles were agents, and which was the object. Objects varied in mass, and agents varied in their relative strength. Each agent always had one personal goal, which could be one of the following: (1) moving itself to a specific landmark, (2) approaching another entity, or (3) moving the object to a specific landmark. In addition to their personal goals, some of the agents also had social goals of either (1) helping the other agent achieve its goal, or (2) hindering the other agent. All animations were 10 seconds long with a framerate of 30.

We generated a large number of stimuli using our hierarchical planner with randomized parameter settings, including (1) entities' sizes and initial positions, (2) the environment layout, (3) agents' strengths, goals, and their relationship. We manually selected 30 representative examples.[7]

Results

The comparison of the cue-based models and the GSPI model is summarized in Table 1. Importantly, human responses were closer to the predictions of the GSPI model than to the ground truth. The bootstrapped inter-subject correlations were $r = .83$ ($SD = .02$) for goals, $r = .88$ ($SD = .03$) for relationships, and $r = .68$ ($SD = .09$) for strengths, which shows that humans made highly consistent inferences of goals and relationships, but less consistent inferences of strengths.

	GSPI	Cue1	Cue2	Cue3	GT
Goal	.84	.06	.09	.09	.73
Relation	.90	.21	.32	.01	.77
Strength	.75	.60	.63	.06	.71

Table 1: Correlations of average human responses with the models and with ground truth.

We calculated goal judgements as the marginalized probability of a goal being reported in a given stimulus. Human responses to the strength question were recorded as (-1, 0, 1), corresponding to ("weaker", "same", and "stronger"). We found that all cue-based models performed well on the training data, but poorly on the testing stimuli. Notably, Cue-based-1 and Cue-based-2 produced good estimates of the agents' relative strength, suggesting that simple cues could be useful in judging physical properties. A detailed summary comparing the GSPI model to human responses is given in Figure 3, showing a close match between the models' and the humans' judgements. As shown in Figure 3D, human and model also agree on which stimuli were easy (low-entropy) and which were hard (high-entropy). The correlation between model and human entropy was $r = .41$ ($p = .02$).

Figure 4 shows four representative stimuli along with the corresponding human and model inferences. The model not only recognized the ground-truth with high confidence in most cases, but also shared similar confusion with humans over goals and relations when the agents' behaviors were hard to interpret (e.g. both humans and the model were all uncertain about the goals and the relationship in Figure 4**E**). Notably, subjects sometimes failed to recognize that an agent also had a personal goal in addition to its social goal (Figure 4**D**). In contrast, in such cases the model generated high confidence inferences over both goals.

Discussion

Our results show that human interpretations of complex social and physical multi-agent interactions can be described by a hierarchical planner combined with a physical simulation engine. Notably, the proposed GSPI model matches human predictions on a number of important dimensions: (1) it can *accurately predict* human responses, even in cases where several interpretations are plausible, (2) it makes *mistakes similar to human mistakes* in cases when ground truth is unclear, and (3) unlike alternative cue-based models, our GSPI model requires *few observations* of behavior to reach an inference. In contrast, cue-based models not only require a large corpus of training data, but are also constrained to more simple interactive scenarios (for example, inferring an agent's strength), and produce less generalizable results.

[6]Stimuli can be viewed at `https://www.youtube.com/playlist?list=PL0ygI9h8RqG_yypVml0xMl8Lkcd5hbuxk`

[7]We aimed to sample a variety of enacted scenarios, and preferred animations that could be interpreted with ambiguity to elicit a distribution of responses over possible interpretations.

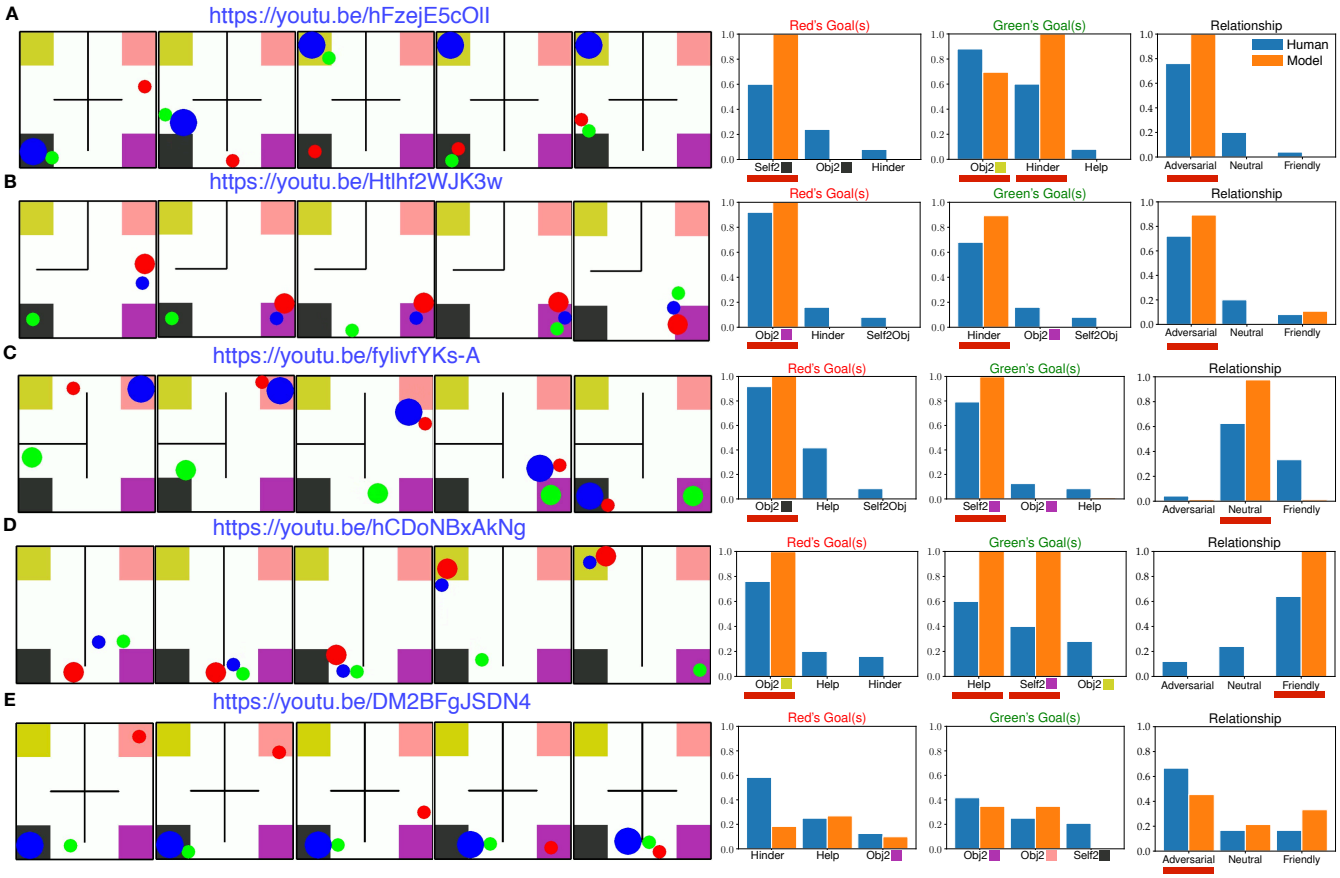

Figure 4: Representative stimuli (left) and their corresponding human and model judgment on goals and relationships (right). The red and green circles are agents and the blue circle is an object. Here, we show the top 3 goals out of all 12 goals for each agent based on human responses. Personal goals are coded as "X2Y", which indicates "get entity X to location Y"; colored squares represent landmarks; "Self" indicates the agent itself; "Obj" means the object entity. Note that the probabilities of goals do not sum up to 1 since an agent could have 1 or 2 goals in a video. The ground-truth is highlighted by red underscore bars (the agents' ground-truth goals in **E** was not among the top 3 human responses.). We include the URL links for viewing the stimuli.

The *Flatland* paradigm offers a convenient, automated, and controllable way of generating a variety of social and physical stimuli. While the present study considered two-agent *Flatland* scenarios with limited goals and properties, the framework can be extended to multiple agents, alternative world layouts, and different physical engines. Together with the GSPI inference engine, *Flatland* improves on the current tools for studying physical inference and social attribution, and allows researchers to study both of these phenomena at the same time.

The model and humans sometimes disagree about the possible agents' goals. Some disagreements occur when the model's confidence is high, but human confidence is low (see the top-left dots in Figure 3A). Individual inspection of the stimuli in question revealed three common cases for disagreement. First, human interpretations of agents' actions are less accurate when agents are weak, leading to noisy estimates of goals. Second, humans sometimes report one of the agent's sub-goal as the final goal, and miss to notice the other goals. Third, humans sometimes fail to recognize the personal goals of a helping or hindering agent. Future work could study sub-goal attribution in more detail, by asking subjects to report all possible goals, along with their probability. Future work could also investigate the richness of human judgements in ambiguous stimuli, by asking subjects to informally describe the reasoning behind their inferences. For example, in highly ambiguous scenarios humans could rely on a library of abstract structures in social situations[8], in order to generate explanations outside the space admissible by our model.

Disagreements may also happen when humans' confidence is high, but the model confidence is low (the bottom-center dots in Figure 3A). Such scenarios are interesting because they may reveal non-uniform priors that humans bring to the table. For example, humans might assume that the agents are friendly or adversarial by default, leading to a biased goal inference. Alternatively, humans might place higher priors on certain types of goals in preference to other types. Such priors may also vary between subjects, with different subjects exhibiting different kinds of non-uniform priors, depending

[8]For example, such abstract social structures could include: jealousy, game-play, sport, flirtation, bluff, disappointment, etc.

on recent experience, context, or personality. We intend to investigate these phenomena in future work.

Lastly, in the current work we assume that both the subjects, and the model, know which entities are agents or objects. This allows us to constrain the inference to goals, relationships, and strengths, for simplicity of analyzing and presenting results. At the same time, telling apart agents from objects (i.e., animacy detection), as humans do easily and intuitively, is an interesting modelling challenge on its own. Judging animacy may require a complex interplay of appearance cues, ability to move on its own, producing the kinds of movement expected of animate agents (e.g. breathing, shivering), as well as the interpretation of actions as intentional and directed toward a goal. Future studies could investigate the mental representations of such inferences, and our experimental paradigm could provide an empirical and computational platform toward supporting this investigation.

A wider implication of our work, is that it demonstrates the computational synergy between intuitive physics, hierarchical planning, and theory of mind. While individual cognitive phenomena in these domains are traditionally considered separate domains, much of cognition likely share this hierarchical, interdependent and multi-sensory structure. This means that inferences informed by different sensory modalities and mental representations produced by different cognitive domains are available to each other. Our work takes a step toward a computational approach of studying the multisensory nature of the mind. We show that it is possible to computationally model how humans interpret complex social and physical scenarios.

Acknowledgement

This work was supported by NSF STC award CCF-1231216, ONR MURI N00014-13-1-0333, and ONR Science of AI grant.

References

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology*, *64*(3), 431.

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*, 0064.

Balaguer, J., Spiers, H., Hassabis, D., & Summerfield, C. (2016). Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron*, *90*(4), 893–903.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Carey, S. (2000). The origin of concepts. *Journal of Cognition and Development*, *1*(1), 37–41.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, *113*(34), E5072–E5081.

Forbus, K. D. (2019). *Qualitative representations: How people reason and learn about the continuous world*. MIT Press.

Gelman, R., Durgin, F., & Kaufman, L. (1995). Distinguishing between animates and inanimates: Not by motion alone. *Causal cognition: A multidisciplinary debate*, 150–184.

Gilden, D. L., & Proffitt, D. R. (1989). Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 372.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Hamlin, K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental science*, *16*(2), 209–226.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 243–259.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103.

Jara-Ettinger, J., Schulz, L., & Tenenbaum, J. B. (2015). The naïve utility calculus: Joint inferences about the costs and rewards of actions. In *Cogsci.*

Kaelbling, L. P., & Lozano-Pérez, T. (2011). Hierarchical task and motion planning in the now. In *Ieee international conference on robotics and automation.*

Kryven, M., Ullman, T., Cowan, W., & Tenenbaum, J. B. (2016). Outcome or strategy? a bayesian model of intelligence attribution. In *Proceedings of the thirty-eighth annual conference of the cognitive science society.*

Michotte, A. (1963). *The perception of causality.* Basic Books.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. *arXiv preprint arXiv:1802.07740*.

Shu, T., Peng, Y., Fan, L., Lu, H., & Zhu, S.-C. (2018). Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations. *Topics in Cognitive Science*, *10*(1), 225–241.

Shu, T., Peng, Y., Lu, H., & Zhu, S.-C. (2019). Partitioning the perception of physical and social events within a unified psychological space. In *Proceedings of the forty-first annual conference of the cognitive science society.*

Sliwa, J., & Freiwald, W. A. (2017). A dedicated network for social interaction processing in the primate brain. *Science*, *356*(6339), 745–749.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2010). Help or hinder: Bayesian models of social goal inference. In *Proceedings of advances in neural information processing systems* (p. 1874-1882).

van Buren, B., Gao, T., & Scholl, B. J. (2017). What are the underlying units of perceived animacy? chasing detection is intrinsically object-based. *Psychonomic bulletin & review*, *24*(5), 1604–1610.

Yildirim, I., Saeed, B., Bennett-Pierre, G., Gerstenberg, T., Tenenbaum, J., & Gweon, H. (2019). Explaining intuitive difficulty judgments by modeling physical effort and risk. *arXiv preprint arXiv:1905.04445*.