PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception

Aviv Netanyahu[*], **Tianmin Shu**[*], **Boris Katz**, **Andrei Barbu**, **Joshua B. Tenenbaum**
Massachusetts Institute of Technology, Cambridge, MA 02139
{avivn, tshu, boris, abarbu, jbt}@mit.edu

Abstract

The ability to perceive and reason about social interactions in the context of physical environments is core to human social intelligence and human-machine cooperation. However, no prior dataset or benchmark has systematically evaluated physically grounded perception of complex social interactions that go beyond short actions, such as high-fiving, or simple group activities, such as gathering. In this work, we create a dataset of physically-grounded abstract social events, PHASE, that resemble a wide range of real-life social interactions by including social concepts such as helping another agent. PHASE consists of 2D animations of pairs of agents moving in a continuous space generated procedurally using a physics engine and a hierarchical planner. Agents have a limited field of view, and can interact with multiple objects, in an environment that has multiple landmarks and obstacles. Using PHASE, we design a social recognition task and a social prediction task. PHASE is validated with human experiments demonstrating that humans perceive rich interactions in the social events, and that the simulated agents behave similarly to humans. As a baseline model, we introduce a Bayesian inverse planning approach, SIMPLE (SIMulation, Planning and Local Estimation), which outperforms state-of-the-art feed-forward neural networks. We hope that PHASE can serve as a difficult new challenge for developing new models that can recognize complex social interactions.

Introduction

Humans make spontaneous and robust judgments of others' mental states (e.g., goals, beliefs, and desires), characteristics (e.g., physical strength), and relationships (e.g., friend, opponent) by watching how other agents interact with the physical world and with each other. These judgements are critical to engaging socially with other agents. AI and robots that cooperate with humans will similarly need to engage with us socially, and by extension make these same judgements about both physical notions, like strength, and social notions, like mental states.

Prior work has looked at recognizing social interactions, but evaluations and benchmarks have been limited to the artifacts of social interactions, like high-fives, hand shakes, or hugging. Here, we create the first benchmark for reasoning about the underlying mechanisms and beliefs of social interactions, instead of these overt actions that are correlated with certain types of interactions. Take a common interaction like helping. It is true that picking up an object next to someone and carrying it in the direction that they are walking towards is likely to be helping, but it might not be. This depends entirely on the intent of the other person and the relationship between these two persons. In the same way that one action does not imply a certain type of social interaction, a certain type of social interaction is also not signaled by a single action. E.g., helping can take a virtually unlimited number of forms. Therefore, we propose a novel dataset, PHASE (PHysically-grounded Abstract Social Events), that

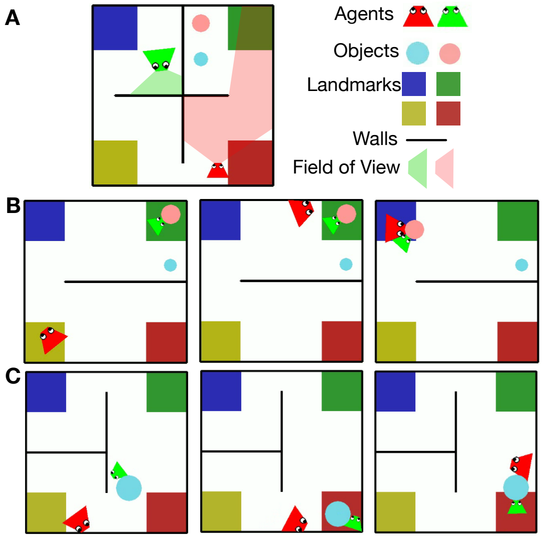

Figure 1: Demonstrating the PHASE physical-social simulation. (A) The elements of the simulation: agents (with a limited conical field of view), objects with different colors and sizes, landmarks with different colors, and immovable walls. (B) Frames from a video depicting an abstract social event sampled in the PHASE simulator. The green agent is weak, therefore has difficulty moving the pink object, which the red agent eventually helps with. (C) Frames from a video depicting the opposite situation, where a green agent is moving an object when the red agent steps in and takes it away.

[*] Equal contribution.

expresses complex social concepts in a physical setting, such as helping and hindering, instead of simple actions.

Collecting datasets of social interactions is very difficult, as it involves playing out complex scenarios while recording the intent and mental states of agents. Heider & Simmel (1944) demonstrated that one can understand social events depicted in animations of simple geometric shapes moving in a physical environment. The movements of these shapes come alive and are interpreted as social behaviors, such as chasing, hiding, or stealing. We draw inspiration from this to build a dataset of physically-grounded abstract social events in a 2D physics engine.

We propose a joint physical-social simulation, as shown in Figure 1, where agents and objects are physical bodies moving in a 2D physics simulation. Agents have a partial observation, are self-propelled allowing them to maneuver in the environment, and can move objects with varying degrees of difficulty depending on their strengths. Agents have different goals and relationships with one another. Grounded social interactions are generated using a hierarchical planner and a physics engine (Figure 2). Manipulating the parameters of this simulation enables us (i) to procedurally generate complex social events that resemble a wide range of real-life social interactions as training data for models, and (ii) to control social and physical variables to create training data with balanced ground-truth labels as well as to carefully design evaluation for generalization in unseen environments and social behaviors.

PHASE consists of 500 video animations depicting diverse social interactions. Each video has a physical configuration (environment layout, physical properties and initial states of agents and objects) and a social configuration (goals of the agents and relationships between agents). Given these configurations, we sample the behaviors of agents with bounded rationality, giving rise to the animated videos. We conduct two human experiments to evaluate the quality of this dataset. The first asks which, if any, social interactions can be recognized by humans in these videos. We find that a diverse set of interactions exist, highlighting the richness of the dataset. The second asks how similar the behavior of the simulated agents is to how humans would behave in the same scenarios — we do this by asking humans to control the agents. We find that there is no significant difference between the human-generated trajectories and the synthetic ones.

We propose two machine social perception tasks on this dataset. The first requires recognizing goals and relationships of agents. The second requires predicting the future trajectories of agents. We test state-of-the-art methods based on feed-forward neural networks and show that they fail to understand or predict many of these social interactions. To further augment machine perception of social interactions, we introduce a Bayesian inverse planning-based approach, SIMPLE (SIMulation, Planning and Local Estimation), that significantly outperforms prior work.

In summary, our contributions include: (i) a joint physical-social simulation for procedurally generating abstract social events grounded in physical environments, (ii) using this engine to generate a first-of-its-kind abstract social events dataset, and (iii) proposing two social perception tasks and a benchmark including state-of-the-art methods and a Bayesian inverse planning-based approach. The dataset and the supplementary material are available at https://www.tshu.io/PHASE.

Related Work

Social Interaction Understanding. Prior work on understanding social interactions has mostly focused on recognizing (i) group activities where multiple people engage in simple activities (e.g., standing in a line, gathering, crossing streets) (Choi and Savarese 2013; Shu et al. 2015; Joo et al. 2015; Alameda-Pineda et al. 2015), (ii) short events in sports activities such as setting a ball in a Volleyball game (Ibrahim et al. 2016), and (iii) human-human interactions such as hugging, kicking, and hand-shaking (Ryoo and Aggarwal 2009; Marszalek, Laptev, and Schmid 2009; Patron-Perez et al. 2012; Yun et al. 2012; Hadfield and Bowden 2013; Van Gemeren, Poppe, and Veltkamp 2016; Shu, Ryoo, and Zhu 2016; Kay et al. 2017; Gu et al. 2018; Monfort et al. 2019; Zhao et al. 2019). The videos in existing datasets for these domains usually only last for a few seconds, and actions of interacting people vary only to a small extent. In contrast, the abstract social events in PHASE depict a wide range of longer and more complex social interactions in a dynamic physical environment, where agents frequently change their motion in order to achieve their goals. In addition, we focus on theory-based inference (recognizing goals and relationships) and not on activity classification.

Human Trajectory Prediction. Our trajectory prediction task is closely related to recent work on pedestrian trajectory prediction (Kitani et al. 2012) which has focued on single agent intent prediction (Xie et al. 2017) and socially appropriate navigation in a crowd (Alahi et al. 2016; Gupta et al. 2018; Huang et al. 2019). Unlike typical pedestrians' movements, agents in our simulation engage in various interactions with each other and with the physical environment, posing a more challenging prediction problem.

Agent-based Social Interaction Simulation. Agent-based simulation has been used for modeling human behavior for a long time (Bonabeau 2002; Railsback, Lytinen, and Jackson 2006). Specifically, to synthesize social interactions for social perception problems, there has been work on using manually defined heuristics to simulate motion trajectories (Gao, McCarthy, and Scholl 2010; Gao and Scholl 2011; Kim et al. 2020), utilizing planning or deep reinforcement learning to generate goal-directed action sequences in simple 2D grid worlds (Ullman et al. 2009; Rabinowitz et al. 2018), and recruiting humans to create animations by moving shapes without physics (Gordon and Roemmele 2014). Recently, there has been work on human social perception attempting to simulate agent behaviors with a physics engine in a fully observable environment for a handful of scenarios (Shu et al. 2019, 2020). Our work extends the idea of simulating agents with a physics engine, but adopts a more complex and realistic setting (e.g., partial observability) and a more sophisticated generative model for simulating richer social behaviors.

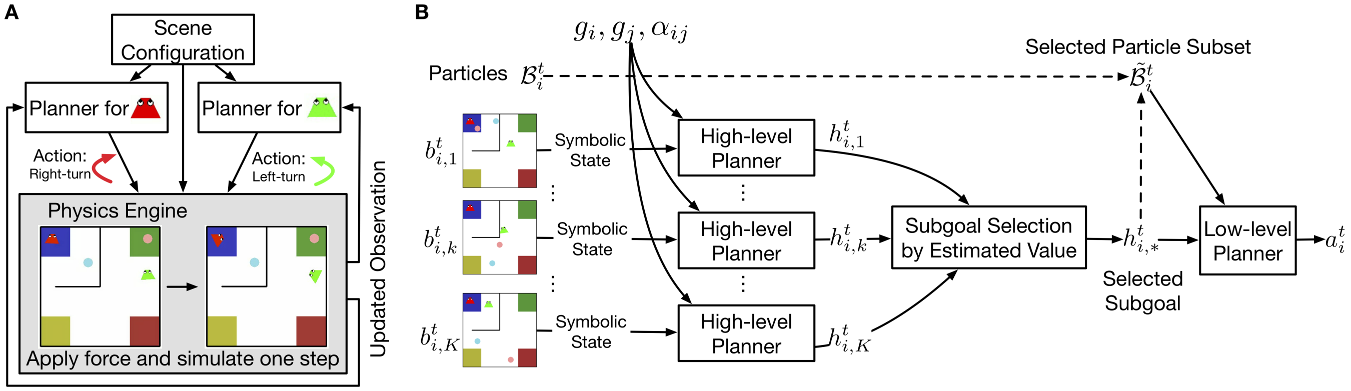

Figure 2: Overview of the simulation and the hierarchical planner. (A) Key components of the simulation. (B) The hierarchical planner in our simulation. At each step the planner searches for an action based on the agent's belief represented by a set of particles. The dashed lines indicate particle subset selection based on the best subgoal.

Synthetic Datasets for Machine Perception. There has been a long tradition on using synthetic data for machine perception (Zitnick, Vedantam, and Parikh 2014; Ros et al. 2016; Johnson et al. 2017; Song et al. 2017; Jiang et al. 2018; Yi et al. 2019; Nan et al. 2020; Cao et al. 2020), since synthetic datasets can be scaled up inexpensively, offer high quality ground-truth annotations without human efforts, and allow for careful experimental control. However, existing synthetic datasets focus on physical scenes (Ros et al. 2016; Song et al. 2017; Johnson et al. 2017) or single agent activities (Nan et al. 2020; Cao et al. 2020). To our knowledge, our PHASE dataset is the first synthetic dataset simulating complex social interactions in dynamic physical environments.

Joint Physical-Social Simulation

The objective of this simulation is to synthesize motion trajectories of multiple entities (agents and objects) that not only follow physical dynamics, but also elicit strong impression of social behaviors. As shown in Figure 2A, the simulation has three main components: a structured physical and social scene configuration, a hierarchical planner, and a physics engine. To synthesize an abstract social event, we first specify the physical configuration, which includes the environment layout, sizes of entities, agents' strengths (maximum forces they can exert), the initial positions of entities, and the social configuration (agents' goals and relationships). Each agent has an independent hierarchical planner that has access to its own mental state and partial view of the environment. At each step, each agent replans based on its beliefs and observations of the scene, and informs the physics engine which action to be applied to its body. The physics engine, then steps the environment forward, resolving the motions of all of the objects. This process repeats to generate a video.

Formulation

We formally define the social behaviors of agents by a decentralized partially observable Markov decision process (Dec-POMDP) (Nair et al. 2003). There are N agents sharing the same state space $\mathcal{S}$ and action space $\mathcal{A}$. In our simulation, the action space consists of applying a force in one of 8 directions, turning right or left, stopping, grabbing an object (attaching it to the agent's body) or letting go of an object, and no force. The physical dynamics of the environment is defined by state transition probabilities $\mathcal{T} : \mathcal{S} \times \mathcal{A}^N \times \mathbb{R}^N \to \mathcal{S}$, i.e., $P(s'|s, \{a_i\}_{i=1}^N, \{f_i\}_{i=1}^N)$, where $f_i \in \mathbb{R}$ is the maximum magnitude of the force agent i can exert at one step, defining the agent's physical strength.

At each step t, agent i observes part of the world state s^t through both vision (which is limited to a conical field of view that is obstructed by internal walls and other entities) and touch (a sensor attached to the body of the agent that reports one of two states, touching or not touching), i.e., $o_i^t \sim O_i(o|s^t)$. The agent updates its belief, $b(s^t)$, based on the current observation by $b(s^{t+1}) \propto O_i(o|s^{t+1}) \sum_{s^t \in \mathcal{S}} P(s^{t+1}|s^t, \{a_i\}_{i=1}^N, \{f_i\}_{i=1}^N) b(s^t)$. All agents know the underlying map of the environment and the total number of agents and objects, but do not know where these other entities are unless they are seen or felt.

Each agent has a physical goal $g_i \in \mathcal{G}$ or a social goal, i.e., helping or hindering. Social goals are indicated by a social utility weight $\alpha_{ij} \in \{-1, 0, 1\}$. When $\alpha_{ij} = 1$, agent i will help agent j achieve its goal; when $\alpha_{ij} = -1$, agent i will hinder agent j; when $\alpha_{ij} = 0$, agent i will pursue its own physical goal. According to this definition, we can write an agent's reward in the context of a 2-agent interaction as

$$R_i(s, a) = (1 - |\alpha_{ij}|)R(s, g_i) + \alpha_{ij} R(s, g_j) + C(a), \quad (1)$$

where $C(a)$ is the cost of taking action a. We choose this definition as similar reward formulation has been shown to be effective for modeling utilities of agents who have different relationships with each other (Kleiman-Weiner, Saxe, and Tenenbaum 2017). Given this reward function, each agent plans its action to maximize accumulated reward over a limited horizon T, i.e., $\sum_{t=1}^T R_i(s^t, a_i^t)$. We assume agents know each other's goals. As shown by our human experiments, this setting can generate rich social behaviors without the additional complexity introduced by the uncertainty of other agents' intentions and strengths. Compared to

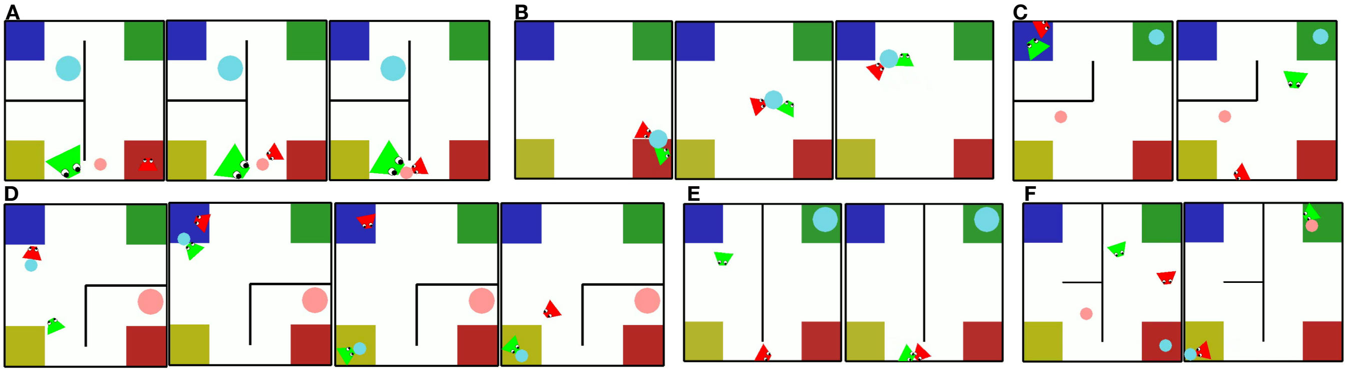

Figure 3: Example abstract social events in the PHASE dataset. (A) Helping a large-sized agent get an object it could not reach. (B) Two weak agents carrying an object together (collaboration). (C) Green chasing red. (D) Two agents trying to put the same object to different landmarks. (E) Red blocking green (hindering). (F) Neutral agents pursuing independent goals.

alternative frameworks such as I-POMDP (Gmytrasiewicz and Doshi 2005), we find that Dec-POMDP offers a good balance between the richness of the resulting agent behaviors and the computational tractability.

Hierarchical Planner

It is challenging to synthesize complex social behaviors at scale. Prior work attempted to do this by manually designing motion heuristics for specific interactions, e.g., chasing (Gao and Scholl 2011). Recent work on deep reinforcement learning demonstrated promising results (Baker et al. 2019), but the trained policies are constrained to a very small goal space and cannot generalize to more diverse situations without re-training. In this work, we propose a hierarchical planner as shown in Figure 2B for deriving agent behaviors with bounded rationality, which is inspired by task and motion planning (TAMP) (Kaelbling and Lozano-Pérez 2011), a framework for solving long-horizon motion planning problems. Note that the main focus of this work is not developing a general-purpose multi-agent planner, therefore the modular design of our simulator allows for the deployment of other planners. We outline the hierarchical planner below and provide more details in the supplementary material.

The planner maintains a set of particles to approximate the belief of each agent at each step, i.e., $\mathcal{B}_i^t = \{b_{i,k}^t\}_{k=1}^K$, where each particle $b_{i,k}^t$ represents a possible world state. All particles are initially sampled from a uniform distribution of possible states of entities. At each step, we first update particles by simulating one step in the physics engine assuming that other agents will maintain a constant motion and then resample the particles that violate the new observation.

Given the current particle set, we use a high-level planner to generate subgoals. The high-level planner first converts the physical state in each particle into symbolic states represented by predicates, and then searches for the best symbolic plan based on the reward of each agent. In particular, we define four types of predicates, On(*agent/object*, *landmark*), Touch(*agent*, *agent/object)*, Attach(*agent*, *object*), Close(*agent/object*, *agent/object/landmark*), and their negations. Subgoals are represented by predicates indicating which immediate states an agent should reach in order to achieve the final goal. This produces a subgoal space $\mathcal{H}$ consisting of all possible predicates. For computational efficiency, we only consider the most immediate subgoal in the plan for the next move. Let $h_{i,k}^t \in \mathcal{H}$ be the best subgoal for agent i at step t based on its belief state in particle $b_{i,k}^t$. We estimate the value of each subgoal by $V(\mathcal{B}_i^t, h, g_i, g_j, \alpha_{ij}) = 1/K \sum_{k=1}^K \mathbb{1}(h = h_{i,k}^t) - \lambda/(\sum_{k=1}^K \mathbb{1}(h = h_{i,k}^t)) \sum_{k=1}^K \mathbb{1}(h = h_{i,k}^t)\hat{C}(b_{i,k}^t, s_g)$, where $\hat{C}(b_{i,k}^t, s_g))$ is a heuristics-based estimation of cost to reach goal state s_g based on belief state $b_{i,k}^t$ defined as the estimated distance that the agent needs to travel before reaching the final goal state, and $\lambda \in (0, 1)$ is a scaling factor. The high-level planner will select the most valuable subgoal at the current step, i.e., $h_{i,*}^t = \arg\max_h V(\mathcal{B}_i^t, h, g_i, g_j, \alpha_{ij})$. Intuitively, this favors a subgoal that frequently appears in the subgoal plans among the particles, and has a lower cost. We illustrate the effect of this value function in the supplementary material.

Finally, we feed the subset of the particles that yield $h_{i,*}^t$ as the best subgoal ($\tilde{\mathcal{B}}_i^t \subset \mathcal{B}_i^t$) to the low-level planner, which will search for the best action to reach that subgoal. In practice, we use A* for the high-level planner, and POMCP (Silver and Veness 2010) for the low-level planner.

PHASE Dataset

The simulator and planner above were used to create the PHASE dataset, the statistics of which are reported below. This dataset was also validated with human experiments.

Procedural Generation

To synthesize the PHASE dataset, we sample a rich set of scene configurations, each of which is fed to the simulation to render a video depicting an abstract social event. In particular, we sample the following variables:

Physical Variables. There are 90 different environment layouts, comprised of wall positions and sizes. There are four possible sizes for entities and four agent strength levels. There are always exactly two agents, and up to two objects.

Social Variables. We sample either a physical goal or a social goal for each agent. The physical goals are: going to one of the four landmarks, moving a specific object to one of the four landmarks, approaching another agent, and getting away from another agent. As there could be two different objects, we have 14 physical goals in total. There are two social goals — helping and hindering.

By sampling the environment layout, entity sizes, agent strengths, agent goals, α_{ij} and α_{ji}, and the initial states of all entities, we can create a large set of physical and social scene configurations. In general, there are five types of social events that appear in the resulting videos: (i) helping an agent overcome an obstacle, (ii) collaborating on a joint goal, (iii) hindering an agent, (iv) two agents having conflicting goals such as chasing or trying to move the same object to different landmarks, and (v) two neutral agents pursuing independent goals. We show representative examples of these social events in Figure 3 and in the supplementary material. Finally, we define three types of relationships based on these five types of social events: (i) and (ii) correspond to friendly relations, (iii) and (iv) correspond to adversarial relations, and (v) corresponds to neutral relations.

Dataset Statistics

PHASE contains 500 videos of abstract social events. Each lasts from 10 sec to 25 sec. Each goal has 47 to 129 examples. For the friendly, adversarial, and neutral relations, there are 200, 192, and 108 examples respectively. With these 500 videos, we create a training set of 320 videos, a validation set of 80 videos, and a testing set of 100 videos. To evaluate the generalization of a trained model, 80% of the testing videos are synthesized with novel environment layouts that are unseen in the training and validation sets. Moreover, there are 9 videos showing unique types of social interactions that are only seen in the test set (e.g., two agents working together towards a common object-related goal).

Human Experiments

To evaluate whether PHASE depicts social interactions, we conduct two human experiments on Mechanical Turk. In both experiments, subjects gave informed consent. The study was approved by the MIT Institutional Review Board.

Experiment 1: Multi-label descriptions. We compiled a set of 23 social interaction types from two sources: (i) common social interactions studied in prior literature (Gao and Scholl 2011; Gordon and Roemmele 2014), and (ii) free responses collected from a preliminary Mechanical Turk experiment where we asked participants to describe a sample set of videos using their own words. We recruited 130 participants to label 20% of the videos in PHASE. Each participant was asked to watch a video and select which of the 23 types of interactions was depicted in the video. In total, each video was judged by 10 participants. We found that all 23 types of interactions were selected to describe at least one video. With a stricter test, measuring if at least half of the participants assigned a particular label to a video, we found that the abstract social events in PHASE resemble 18 diverse real-life interaction categories (Figure 4), participants could recognize unintentional interactions (e.g., when agents with independent goals accidentally crossed paths), and all friendly and adversarial interactions were meaningful and intentional to participants.

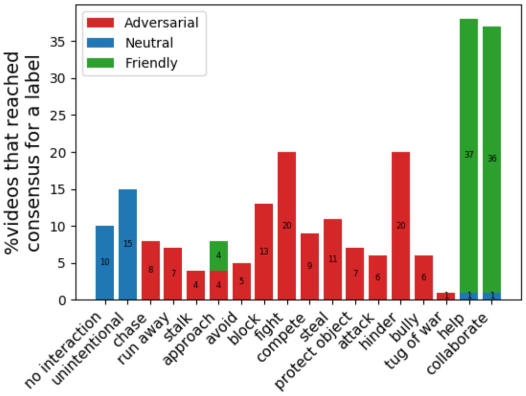

Figure 4: Consistent human responses in Experiment 2 showing how many videos (percentages) were assigned with an interaction category by at least 50% of the participants who have watched the videos.

Experiment 2: Comparing the synthesized trajectories with human-controlled trajectories. This experiment consists of two parts. In the first part, we designed a 2-player game based on PHASE, where humans can control agents by pressing keys. Using interface, we collected 100 videos by asking three human controllers to play with each other in the scenarios that were matched with the scene configurations of 100 test videos in PHASE.

In the second part, we recruited 186 additional participants on Mechanical Turk and divided them into two groups. One group watched the human-controlled videos, and the other watched matching videos from PHASE. For each video, participants were asked to judge the goals and relations of the agents, and rate how likely humans were to behave similarly to these agents under the same goals and relations (on a scale of 1 to 5). In both groups, participants achieved a high accuracy for goal and relation recognition (0.965 and 0.92 for goal and relation recognition on the human-controlled videos, and 0.97 and 0.99 for goal and relation recognition on the PHASE videos). The averaged human-likelihood rating for the human-controlled videos is 4.06 ($\sigma = 0.36$); and for PHASE, it is 3.98 ($\sigma = 0.42$). This suggests that to the participants, (i) the PHASE videos and the human-controlled videos exhibit similar social events in terms of goals and relationships, even though they have different motion trajectories, and (ii) the agent behaviors in these two types of videos all have similar degrees of human-likelihood.

Social Perception Tasks

We design two social perception tasks that evaluate a model's abilities to recognize the goals and relations of agents, and to predict the future social behaviors.

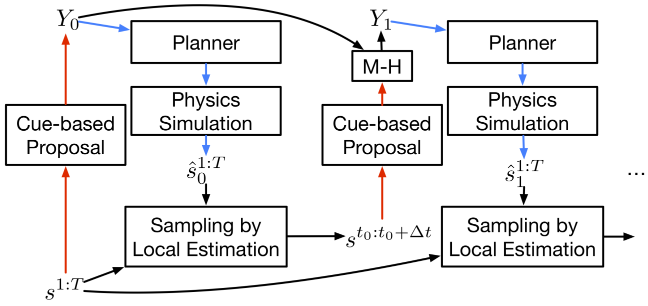

Figure 5: Diagram of how the proposal in a single particle is updated in SIMPLE. For brevity, we drop the subscript m. M-H represents Metropolis–Hastings algorithm for determining whether to accept the new proposal. The red lines indicate bottom-up proposals and the blue lines represent top-down generative processes.

Task 1: Joint inference of goals and relations. This task focuses on understanding social interactions, i.e., jointly inferring agents' goals and relationships to other agents to explain their behavior. Unlike typical activity recognition, this task requires a more fundamental understanding of social interactions — we focus on *why* the agents exhibit certain behaviors, rather than giving a literal description of *what* the agents are doing.

Task 2: Multi-entity trajectory prediction. Since robots and intelligent machines must not only understand social interactions, but also engage with us socially, we design a second task to predict the behavior of a social agent. This requires both social and physical reasoning, as all agents and objects are constrained by physics. A model must predict the trajectories in the next 20 steps (5 sec) of all agents after watching the first 20 steps (5 sec).

Inference by Bayesian Inverse Planning

To address the proposed tasks, we develop a Bayesian inverse planning approach, SIMPLE (SIMulation, Planning and Local Estimation), that integrates computational theory of mind (Baker et al. 2017) with simulation for physical reasoning (Battaglia, Hamrick, and Tenenbaum 2013).

Inverse planning relies on the notion that if one correctly infers hidden variables, like goals, relations, strengths, and beliefs, the generated rational plans for agents will be a good match for the observed plans. We instantiate this idea in the following way: let $Y = \langle g_i, g_j, \alpha_{ij}, \alpha_{ji}, f_i, f_j \rangle$ be the hypothesis, $s^{1:T}$ be the observed state sequence (in particular, trajectories of all entities) of the event, and $\hat{s}^{1:T} = G(g_i, g_j, \alpha_{ij}, \alpha_{ji}, f_i, f_j)$ be the simulation given the hypothesis, where the generative model $G(\cdot)$ includes both the hierarchical planner and the physics engine. Then we have the following posterior probability for inference

$$\begin{aligned} &P(Y = \langle g_i, g_j, \alpha_{ij}, \alpha_{ji}, f_i, f_j \rangle | s^{1:T}) \\ &\propto P(s^{1:T}|Y) P(g_i) P(g_j) P(\alpha_{ij}, \alpha_{ji}) P(f_i) P(f_j), \end{aligned} \tag{2}$$

where $P(s^{1:T}|Y) = e^{-\beta \sum_{t=1}^{T} ||s^t - \hat{s}^t||_2}$ is the likelihood based on the distance between the observed trajectories and the simulated trajectories w.r.t. the hypothesis, and $\beta > 0$ is a constant coefficient.

To efficiently explore the large hypothesis space, we perform probabilistic inference based on data-driven Markov Chain Monte Carlo (MCMC) that utilizes both cue-based bottom up proposals and top-down generative processes, as shown by Figure 5. We outline inference in two main steps as follows and discuss additional implementation details in the supplementary material.

Initial Proposals. Even though visual cues of trajectories alone may not give us the most accurate inference, they can provide reasonable guesses which may shrink the search space, thereby increasing the chance of making good proposals. Thus we use a bottom-up proposal approach that estimates the likelihood of pursuing a goal by the distance between the final state and the the goal state as well as the change in that distance compared to the start of the video. For the social utility weights, α, we adopt a uniform distribution. For agent strengths, we train a regression model based on a 2-layer MLP which takes in the average, maximum, and minimum velocities as well as accelerations of each agent. We sample M particles to approximate the true posterior probability (Eq. 2), each of which contains an initial proposal $Y_{0,m}$ sampled from a cue-based proposal distribution, $Q(Y|s^{1:T})$.

Proposal Update based on Local Estimation. We run multiple iterations to update the proposals. Given the proposals at iteration l, we simulate the trajectories, i.e., $\hat{s}_{l,m}^{1:T}$, $\forall m = 1, \cdots, M$, and compare them with the observed trajectories, $s^{1:T}$. For each proposal, we sample a time interval with a fixed length, ΔT, based on the errors between the simulation and the observations, i.e., $t_{l,m} \propto e^{\eta \sum_{\tau=t_{l,m}}^{t_{l,m}+\Delta T} ||\hat{s}_{l,m}^{\tau} - s^{\tau}||_2}$, where $\eta = 0.1$. The intuition behind this is that local deviation is often more informative in terms of how the proposal should be updated compared to the overall deviation.[1] After selecting a local time interval, we use the same bottom-up mechanism to again propose a new hypothesis for each particle, Y'_m, based only on $S' = s^{t_{l,m}:t_{l,m}+\Delta T}$. We then use the Metropolis–Hastings algorithm to decide whether to accept this new proposal for the particle, where the acceptance rate is $\alpha = \min\{1, \frac{Q(Y'|S')P(s^{1:T}|Y')}{Q(Y_{l,m}|S')P(s^{1:T}|Y_{l,m})}\}$.

When planning the actions at step t, the planner utilizes the belief inferred from agents' past observation upon t. We achieve this by estimating the observations of agents at each step using the simulator, and then sample belief particles for each agent that are consistent with what that agent has seen. This purely bottom-up belief estimation can adequately approximate the true beliefs of agents while being computationally efficient. In contrast, proposing beliefs top-down would be intractable due to the large state space.

To approximate the posterior probability, we compute

[1] E.g., in hindering interactions, it is often not clear which physical goal was being hindered once two agents made contact; however, the first part of the video may reveal more information about what an agent's physical goal was since the agent was pursuing that goal without interference from the other agent who was far away.

the weight for each particle m at iteration l as $w_{l,m} = P(s^{1:T}|Y_{l,m})/\sum_{k=1}^{M} P(s^{1:T}|Y_{l,k})$. Then an agent's goal can be inferred by

$$P(g_i|s^{1:T}) = \sum_{m=1}^{M} \mathbb{1}(g_i \in Y_{l,m}) w_{l,m}, \qquad (3)$$

where $\mathbb{1}(g_i \in Y_{l,m})$ indicates whether g_i appears in the hypothesis $Y_{l,m}$. Similarly, we can compute $P(\alpha_{ij}|s^{1:T})$ and $P(\alpha_{ji}|s^{1:T})$. Finally, we can define the posterior probabilities of relationships in terms of the goal and social weights inference. Here we show the probability for the friendly relationship as an example (see the supplementary material for the other two relationships):

$$\begin{aligned} P(\text{friendly}|s^{1:T}) = \;& P(\alpha_{ij} > 0 \text{ or } \alpha_{ji} > 0|s^{1:T}) \\ & + P(g_i = g_j|s^{1:T}) \\ & \cdot P(\alpha_{ij} = 0, \alpha_{ji} = 0|s^{1:T}). \end{aligned} \qquad (4)$$

This same model can be used to simulate future trajectories based on the goal and relation inference. Specifically, we simulate future trajectories for the most likely hypothesized goals and relationships inferred from the prior observation.

Results

For the first task, joint goal and relation inference, we compare our model, SIMPLE (with 15 particles and 6 iterations), with two state-of-the-art approaches for recognizing group activities modified for our domain (see details in the supplementary material) as well as a human performance:

2-Level LSTM: A hierarchical LSTM-based model (Ibrahim et al. 2016) for recognizing individual actions and the overall group activity.

ARG: Actor Relation Graph (Wu et al. 2019), a graph neural net modeling human relations and interactions.

Human: We collected human judgments of goals and relations on the testing videos in the second human experiment. We use this result as a human performance.

For the second task, trajectory prediction, in addition to evaluating online trajectory prediction using our hierarchical planner and inference based on SIMPLE, we similarly adopt two feed-forward models as baselines:

Social-LSTM: LSTM-based trajectory prediction with a social pooling mechanism (Alahi et al. 2016).

STGAT: Spatial-Temporal Graph Attention network (Huang et al. 2019), a state-of-the-art multi-person trajectory prediction approach.

We use two metrics common in prior work on trajectory prediction (Alahi et al. 2016): Average Displacement Error (ADE), i.e., average L2 distance between ground truth and the prediction over all steps, and Final Displacement Error (FDE), i.e., the distance between the predicted position and the ground-truth position at the last step. Note that we compute the distance only based on the positions of the entities and do not consider their velocities and angles.

Table 1 and Table 2 summarize the performance of all methods in the two tasks. For the first task, humans achieve almost perfect accuracy. SIMPLE performs significantly better than the other two baselines based on feed-foward neural nets. This suggests that the underlying meaning of different social interactions could not be captured by motion patterns alone. Similarly, the trajectory prediction based on SIMPLE also outperforms both Social-LSTM and STGAT. Moreover, the results also suggest that prediction in videos where two friendly or adversarial agents engage in social interactions is more challenging than prediction in videos of two neutral agents pursing their own independent goals.

Method	Goal			Relation
	Top-1	Top-2	Top-3	
Human	0.970	0.990	0.995	0.99
2-Level LSTM	0.405	0.610	0.735	0.77
ARG	0.485	0.665	0.805	0.61
SIMPLE	0.870	0.890	0.910	0.88

Table 1: Goal and relation recognition accuracy in Task 1.

Method	ADE			FDE		
	Ind.	Social	Ave	Ind.	Social	Ave
S-LSTM	6.47	7.21	7.02	7.10	7.93	7.72
STGAT	6.64	7.23	7.08	7.44	7.79	7.70
SIMPLE	3.39	4.14	3.84	4.42	5.75	5.23

Table 2: Trajectory prediction error in Task 2. We report the prediction errors for independent videos, social (friendly or adversarial) videos, and all videos respectively.

Although SIMPLE demonstrates superior results compared to strong baselines, it requires simulation in a physics engine and expensive search with a planner. On the one hand, this shows that with the right kinds of priors (e.g., physical dynamics, goal-directed rational behaviors) and model structures (e.g., computational theory-of-mind), it is possible to achieve understanding and forecasting of the social interactions simulated in PHASE; on the other hand, our work also poses challenges for future work on machine social perception. E.g., how to achieve efficient theory-based inference and social behavior prediction? How can models generalize social and physical dynamics learned from training environments to novel environments?

Conclusion

We propose a joint physical-social simulation to procedurally generate a large set of social interactions grounded in physical environments. We use this simulator to create the first physically-grounded abstract social event dataset, PHASE. Our human experiments show that the videos which comprise PHASE are recognized as depicting a large variety of real-life social interactions. The two social perception tasks for machines demonstrate that much remains to be done with existing models, even with the Bayesian inverse-planning approach, SIMPLE, we introduce for solving these two tasks. Having a systematic benchmark for understanding social interactions will, we hope, spur new research and new models. In the future, we intend to simulate more sophisticated social interactions that require additional features such as expression of emotion, communication, and/or higher-order theory of mind.

Acknowledgements

We would like to thank David Mayo and Yen-Ling Kuo for their help with the experiments. This work was supported by the Center for Brains, Minds, and Machines, NSF STC award CCF-1231216, ONR MURI N00014-13-1-0333, the United States Air Force Research Laboratory under Cooperative Agreement Number FA8750-19-2-1000, Toyota Research Institute, the DARPA GAILA program, and the ONR Science of Artificial Intelligence program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 961–971.

Alameda-Pineda, X.; Staiano, J.; Subramanian, R.; Batrinca, L.; Ricci, E.; Lepri, B.; Lanz, O.; and Sebe, N. 2015. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE transactions on pattern analysis and machine intelligence* 38(8): 1707–1720.

Baker, B.; Kanitscheider, I.; Markov, T.; Wu, Y.; Powell, G.; McGrew, B.; and Mordatch, I. 2019. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528* .

Baker, C. L.; Jara-Ettinger, J.; Saxe, R.; and Tenenbaum, J. B. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1(4): 1–10.

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

Bonabeau, E. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences* 99(suppl 3): 7280–7287.

Cao, Z.; Gao, H.; Mangalam, K.; Cai, Q.-Z.; Vo, M.; and Malik, J. 2020. Long-term Human Motion Prediction with Scene Context. *arXiv preprint arXiv:2007.03672* .

Choi, W.; and Savarese, S. 2013. Understanding collective activitiesof people from videos. *IEEE transactions on pattern analysis and machine intelligence* 36(6): 1242–1257.

Gao, T.; McCarthy, G.; and Scholl, B. J. 2010. The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological science* 21(12): 1845–1853.

Gao, T.; and Scholl, B. J. 2011. Chasing vs. stalking: interrupting the perception of animacy. *Journal of experimental psychology: Human perception and performance* 37(3): 669.

Gmytrasiewicz, P. J.; and Doshi, P. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research* 24: 49–79.

Gordon, A. S.; and Roemmele, M. 2014. An authoring tool for movies in the style of Heider and Simmel. In *International Conference on Interactive Digital Storytelling*, 49–60. Springer.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6047–6056.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2255–2264.

Hadfield, S.; and Bowden, R. 2013. Hollywood 3D: Recognizing actions in 3D natural scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3398–3405.

Heider, F.; and Simmel, M. 1944. An experimental study of apparent behavior. *The American journal of psychology* 57(2): 243–259.

Huang, Y.; Bi, H.; Li, Z.; Mao, T.; and Wang, Z. 2019. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 6272–6281.

Ibrahim, M. S.; Muralidharan, S.; Deng, Z.; Vahdat, A.; and Mori, G. 2016. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1971–1980.

Jiang, C.; Qi, S.; Zhu, Y.; Huang, S.; Lin, J.; Yu, L.-F.; Terzopoulos, D.; and Zhu, S.-C. 2018. Configurable 3d scene synthesis and 2d image rendering with per-pixel ground truth using stochastic grammars. *International Journal of Computer Vision* 126(9): 920–941.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2901–2910.

Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; and Sheikh, Y. 2015. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, 3334–3342.

Kaelbling, L. P.; and Lozano-Pérez, T. 2011. Hierarchical task and motion planning in the now. In *2011 IEEE International Conference on Robotics and Automation*, 1470–1477. IEEE.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* .

Kim, K. H.; Sano, M.; De Freitas, J.; Haber, N.; and Yamins, D. 2020. Active world model learning in agent-rich environments with progress curiosity. In *International Conference on Machine Learning (ICML)*.

Kitani, K. M.; Ziebart, B. D.; Bagnell, J. A.; and Hebert, M. 2012. Activity forecasting. In *European Conference on Computer Vision*, 201–214. Springer.

Kleiman-Weiner, M.; Saxe, R.; and Tenenbaum, J. B. 2017. Learning a commonsense moral theory. *Cognition* 167: 107–123.

Marszalek, M.; Laptev, I.; and Schmid, C. 2009. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2929–2936. IEEE.

Monfort, M.; Andonian, A.; Zhou, B.; Ramakrishnan, K.; Bargal, S. A.; Yan, T.; Brown, L.; Fan, Q.; Gutfreund, D.; Vondrick, C.; et al. 2019. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence* 42(2): 502–508.

Nair, R.; Tambe, M.; Yokoo, M.; Pynadath, D.; and Marsella, S. 2003. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *IJCAI*, volume 3, 705–711.

Nan, Z.; Shu, T.; Gong, R.; Wang, S.; Wei, P.; Zhu, S.-C.; and Zheng, N. 2020. Learning to infer human attention in daily activities. *Pattern Recognition* 107314.

Patron-Perez, A.; Marszalek, M.; Reid, I.; and Zisserman, A. 2012. Structured learning of human interactions in TV shows. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(12): 2441–2453.

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

Railsback, S. F.; Lytinen, S. L.; and Jackson, S. K. 2006. Agent-based simulation platforms: Review and development recommendations. *Simulation* 82(9): 609–623.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3234–3243.

Ryoo, M. S.; and Aggarwal, J. K. 2009. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In *IEEE International Conference on Computer Vision (ICCV)*.

Shu, T.; Kryven, M.; Ullman, T. D.; and Tenenbaum, J. B. 2020. Adventures in Flatland: Perceiving Social Interactions Under Physical Dynamics. In *42nd Annual Meeting of the Cognitive Science Society (CogSci)*.

Shu, T.; Peng, Y.; Lu, H.; and Zhu, S.-C. 2019. Partitioning the Perception of Physical and Social Events Within a Unified Psychological Space. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci)*.

Shu, T.; Ryoo, M. S.; and Zhu, S.-C. 2016. Learning Social Affordance for Human-Robot Interaction. In *International Joint Conference on Aritifical Inteliigence (IJCAI)*.

Shu, T.; Xie, D.; Rothrock, B.; Todorovic, S.; and Chun Zhu, S. 2015. Joint inference of groups, events and human roles in aerial videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4576–4584.

Silver, D.; and Veness, J. 2010. Monte-Carlo planning in large POMDPs. In *Advances in neural information processing systems*, 2164–2172.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1746–1754.

Ullman, T.; Baker, C.; Macindoe, O.; Evans, O.; Goodman, N.; and Tenenbaum, J. B. 2009. Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems*, 1874–1882.

Van Gemeren, C.; Poppe, R.; and Veltkamp, R. C. 2016. Spatio-temporal detection of fine-grained dyadic human interactions. In *International Workshop on Human Behavior Understanding*, 116–133. Springer.

Wu, J.; Wang, L.; Wang, L.; Guo, J.; and Wu, G. 2019. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9964–9974.

Xie, D.; Shu, T.; Todorovic, S.; and Zhu, S.-C. 2017. Learning and inferring "dark matter" and predicting human intents and trajectories in videos. *IEEE transactions on pattern analysis and machine intelligence* 40(7): 1639–1652.

Yi, K.; Gan, C.; Li, Y.; Kohli, P.; Wu, J.; Torralba, A.; and Tenenbaum, J. B. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442* .

Yun, K.; Honorio, J.; Chattopadhyay, D.; Berg, T. L.; and Samaras, D. 2012. Two-person Interaction Detection Using Body-Pose Features and Multiple Instance Learning. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE.

Zhao, H.; Torralba, A.; Torresani, L.; and Yan, Z. 2019. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 8668–8678.

Zitnick, C. L.; Vedantam, R.; and Parikh, D. 2014. Adopting abstract images for semantic scene understanding. *IEEE transactions on pattern analysis and machine intelligence* 38(4): 627–638.