Representing physical and social events within a unified psychological space

Tianmin Shu[1,*], Yujia Peng[3,6], Song-Chun Zhu[4,5,6], and Hongjing Lu[2]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA
[2]Department of Psychology, University of California, Los Angeles, USA
[3]School of Psychological and Cognitive Sciences, Peking University, China
[4]Beijing Institute for General Artificial Intelligence, China
[5]Department of Automation, Tsinghua University, China
[6]Institute for Artificial Intelligence, Peking University, China

Abstract

One of the great feats of human perception is the generation of quick impressions of both physical and social events based on sparse displays of motion trajectories. Here we aim to provide a unified theory that captures the interconnections between perception of physical and social events. A simulation-based approach is used to generate a variety of animations depicting rich behavioral patterns. Human experiments used these animations to reveal that perception of dynamic stimuli undergoes a gradual transition from physical to social events. A learning-based computational framework is proposed to account for human judgments. The model learns to identify latent forces by inferring a family of potential functions capturing physical laws, and value functions describing the goals of agents. The model projects new animations into a sociophysical space with two psychological dimensions: an intuitive sense of whether physical laws are violated, and an impression of whether an agent possesses intentions to perform goal-directed actions. This derived sociophysical space predicts a meaningful partition between physical and social events, as well as a gradual transition from physical to social perception. The space also predicts human judgments of whether individual objects are lifeless objects in motion, or human agents performing goal-directed actions. These results demonstrate that a theoretical unification based on physical potential functions and goal-related values can account for the human ability to form an immediate impression of physical and social events. This ability provides an important pathway from perception to higher cognition.

***Keywords*—** social perception; intuitive physics; intention; deep reinforcement learning; Heider-Simmel animations

1 Introduction

Humans are social animals living in a physical world. We spontaneously endow the world with both physical and social meaning. Every glance into our visual world captures a scene in which inanimate physical and intentional social events closely interact with each other. For example, a quick look out of your window might reveal an eventful scene in which multiple people and objects are in motion. Moreover, these fundamental perceptual abilities operate very rapidly (in a

*To whom correspondence should be addressed. E-mail: tshu@mit.edu. The majority of this work was done while the authors were all working at UCLA.

fraction of second) on impoverished visual stimuli (Scholl & Tremoulet, 2000), are present even in infancy (Leslie & Keeble, 1987), and appear to depend neither on explicit reasoning about causes (Danks, 2009) nor on a sophisticated theory of mind (Burge, 2018).

In psychological research, two classic types of simple animations have been shown to elicit either a vivid impression of physical causation, or of social attribution. Michotte (1963) showed that physical causality can be directly perceived from a simple animation depicting a moving ball colliding with a stationary ball, which then appears to launch and move off. Two decades earlier, Heider & Simmel (1944) had created a film showing the movements of simple geometric shapes: two triangles and a circle moving in the vicinity of a rectangle. These researchers found that people are predisposed to describe such simple shapes as persons with well-defined intentions and social traits. Building on these classic studies, recent psychological research (e.g., Kassin, 1981; Scholl & Tremoulet, 2000; Battaglia et al., 2013) including infancy studies (e.g., Leslie & Keeble, 1987; Csibra et al., 1999; Gergely et al., 1995; Liu et al., 2017), suggest that the ability to ground physical and social perception in simple visual displays may play a key role in the origin of human intelligence. Indeed, Michotte (1963) suggested that this ability provides the foundation for human cognition, serving as an important pathway from perception to reasoning.

Yet despite decades of research on both intuitive physics and social perception, with studies in both areas focusing on dynamic displays consisting of simple shapes in motion, the two research areas have proceeded along parallel tracks. Modeling work in intuitive physics has considered whether humans use heuristics or mental simulation to make sense of physical events (Battaglia et al., 2013; Kubricht et al., 2017). In contrast, theoretical work in social perception has aimed to identify critical motion cues that influence the perception of animacy (Dittrich & Lea, 1994a; Scholl & Tremoulet, 2000; Gao et al., 2009; Shu et al., 2018), and the generation of inferences about theory of mind to interpret goal-directed actions (Baker et al., 2009, 2017; Ullman et al., 2010; Pantelis et al., 2014). More recently, there has also been work on training neural networks using synthetic social events to account for human social perception (Hovaidi-Ardestani et al., 2018). Rather than taking this "divide and conquer" strategy to develop different sets of models for each domain, the present work explores a unified modeling approach to capture physical and social regimes as a continuous spectrum in the model landscape.

Our approach is motivated by the view that perception of both physical and social events is guided by physical and social knowledge that facilitates detection of motion features congruent with it (Dittrich & Lea, 1994b; Gelman et al., 1995). We posit that human judgments are driven by an effort to make intuitive sense of the forces that determine motion of entities. Some forces arise from physical interactions (e.g., collision), whereas other forces reflect the efforts of free-willed agents to pursue their goals (e.g., to attack a person or to move an object). It is possible, for example, that an object first seen as a leaf blowing in the wind will be reinterpreted as a moth, when it appears to evade an attempt to capture it. In that moment, an interpretation of physical forces no longer applies and is replaced by an intentional interpretation based on animacy. Our approach assumes that the physical and social knowledge underlying forces is not coded in terms of prespecified physical laws and social rules, but rather can be learned from observational data in a probabilistic manner.

Here we propose that human judgments are driven by intuitive inference of *physical-social forces* (PSF) as shown in Figure 1, based on identifying relevant perceptual variables and learning functional relations to infer hidden forces. Within the PSF model, the computational goal is to derive the forces governing the movements of entities in both physical and social events. The central innovation is to derive inferences about two key sets of functional relations among entities and the environment: *potential functions* that capture physical laws and interactions, and *value functions* that describe the goals of human agents. Both components can be learned from a small set of observations. The trained model can derive physical-social forces for an animation, and quantify indices for two fundamental dimensions that form a unified space: *deviation from predictions of physical model*, i.e., an intuitive sense of whether physical laws are obeyed or violated; and *degree of goal-directed intention*, i.e., an impression of whether goal-directed behaviors are exhibited. This unified space illuminates how physical and social events can be distinguished, and also how they can be perceived as a continuum.

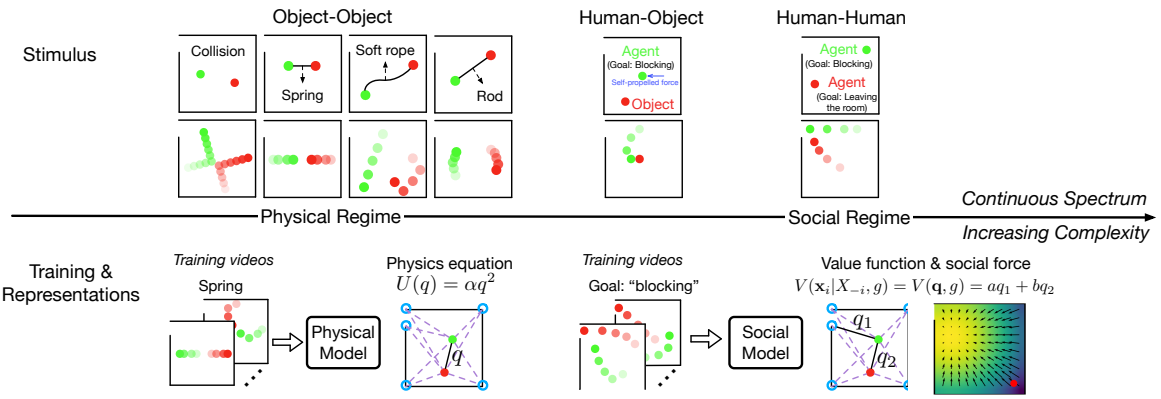


Figure 1: Overview of the PSF model. The PSF model takes inputs of motion trajectories from simple shapes involved in a range of events from physical to social regimes. From one domain to another, our approach learns a family of physical and social models in isolation, augmenting the potential energy terms as it sees more and more events in different domain. Using a small sets of training examples of physical events, the PSF model learns to select a minimum set of perceptual variables as generalized coordinates (highlighted in black line segments) from a pool of candidate coordinates (depicted as purple dashed lines for a subset of candidates), and infers appropriate physical potential functions to capture physical knowledge. Using the other set of training examples of social events, the PSF model selects perceptual variables relevant to social relations and learns value functions encapsulating agents' optimal plans for different goals. Here, green and red circles represent the two entities. The solid line segments represent the selected components $\mathbf{q}$, as annotated in the figure and shown in the terms of potential functions. The dashed line segments represent the remaining component candidates that were not selected by the learning algorithm. value function is depicted as a heat-map in which the color transits from blue to yellow indicating an increase in the value of a state. This illustration of the learned value function reveals that the best action plan to block an agent is to stand between the door and that other agent who aims to leave the room (presented as the red dot in the force field), as this location corresponds to the state with the highest value.

We first tested the model predictions for five classic stimuli used to study causal perception in the literature. But in order to experimentally test whether the model-derived space can account for the perception of physical and social events, we need a large number of animations in which movements of simple shapes reflect a variety of complex events in both physical and social domains. Prior work has typically created such stimuli using manually-designed interactions (Gao et al., 2009, 2010; Gordon & Roemmele, 2014; Isik et al., 2017), simulations of rule-based behavior (Kerr & Cohen, 2010; Pantelis et al., 2014; Sano et al., 2020), or trajectories extracted from human activities in aerial videos (Shu et al., 2018). However, these methods of stimulus generation are unable to produce a large set of animations depicting rich behaviors and showing violations of physical and social constraints in a continuous and controlled manner. Accordingly, we developed a simulation-based approach by integrating a physics engine with deep reinforcement learning to generate hundreds of physical/social animations for use in psychological experiments. This large set of stimuli made it possible to quantitatively assess the proposed model and the hypothesized common psychological space embedding perception of both physical and social events. In two experiments, we show that the derived sociophysics psychological space can provide an effective representation system. The model can predict a meaningful partition between physical and social events that is consistent with human judgments. In addition, the relative locations of individual entities in the sociophysics space can be used to assign agent/patient roles to individual entities (e.g., as a human agent or as an inanimate object).

2 Computational model of physical and social forces

Three key design elements are introduced into the PSF model. First, the PSF adopts a statistical framework to cope with the uncertainty of force sources in physical and social scenes. Second, the PSF employs an efficient representation system to derive hidden forces by selecting optimal perceptual variables, inferring potential functions (U) that capture physical constrains, and estimating value functions (V) that reflect the goals of agents. The force-based representations enable placing physical and social events on a continuous spectrum by incorporating more perceptual variables, and elaborating potential/value functions in more complex situations. Third, the PSF is not provided with deterministic physical laws or social rules; rather, it uses a relatively small number of training examples to learn physical and social knowledge. Specifically, the PSF model learns to discover appropriate perceptual variables and latent functions to derive forces that provide good predictions of dynamic movements in physical and social events.

In our study, the PSF model adopts Lagrangian representations to infer force fields that govern the dynamics of both physical and social events. Figure 1 illustrates the PSF model architecture. Using a fixed set of candidate generalized coordinates as possible perceptual variables (e.g., distance between two entities, distances to environment landmarks), we developed a learning algorithm to select the best variables from this candidate pool. For the potential energy functions, the learning algorithm estimates the parameters for polynomial functions based on the training data. After learning from some examples of physical events, the model uses selected generalized coordinates and potential functions to derive force fields, which can be used to make predictions of entity movements in a dynamic stimulus. The deviation between predicted movements and observed motion is used to revise the selection of generalized coordinates and estimates of potential functions.

2.1 Lagrangian force representation

Formally, in an N-entity dynamic system, we observe the Cartesian coordinates from the positions of individual entities, $X = \{\mathbf{x}_i\}_{i=1}^{N}$ over time. The PSF model takes the inputs to derive forces that govern the movements of entities, akin to force-based analysis in physics. To capture the dynamics of a complex system, we adopted an approach commonly used in physics, Lagrangian mechanics. An alternative representation framework using Cartesian coordinates requires explicit inclusion of constraint forces, which makes force inference difficult for complex systems. The key idea for the force representations in the PSF model is to convert Cartesian coordinates for entities into a generalized coordinate system, $\mathbf{q} = (q_j)_{j=1}^{D}$, where D indicates the degrees of freedom in the dynamic system, and then define potential energy functions of generalized coordinates to derive forces applied to individual entities in the system. Each dimension in the generalized coordinates, q_j, is derived by a transformation function $q_j = q_j(\mathbf{x}_1, \cdots, \mathbf{x}_N)$. For example, a system consisting of two objects that are connected with a spring (Figure 2a) can be conveniently defined by only one variable – the distance between the two entities as generalized coordinates, i.e., $q = q(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{x}_1 - \mathbf{x}_2||$. The use of generalized coordinates q and potential functions $U(\mathbf{q})$ allows parsimonious computation to derive the force F applied to the entity in the location $\mathbf{x}$ by:

$$\hat{\mathbf{F}}_i = -\sum_{j=1}^{D} \frac{\partial U(\mathbf{q})}{\partial q_j} \frac{\partial q_j}{\partial \mathbf{x}_i} \quad \forall i = 1, \cdots, N. \tag{1}$$

Lagrangian force representation allows multiple potential energy functions to coexist in a complex dynamic system. Based on the additive property of the Lagrangian representations, the overall potential energy is simply the sum of all individual potential energy functions, i.e., $U(\mathbf{q}) = \sum_{j=1}^{D} U_j(q_j)$. For instance, in Figure 2b, by defining generalized coordinates $q_1 = ||\mathbf{x}_1 - \mathbf{x}_2||$ and $q_2 = ||\mathbf{x}_1 - \mathbf{x}_3||$, the overall potential energy for this two-spring system can be captured by adding up two functions associated with individual springs: $U(\mathbf{q}) = U_1(q_1) + U_2(q_2)$. If the two springs have the same property, then the potential energy can be further simplified by reusing the same atomic function: $U(\mathbf{q}) = U(q_1) + U(q_2)$. In this work, we assume a polynomial form for

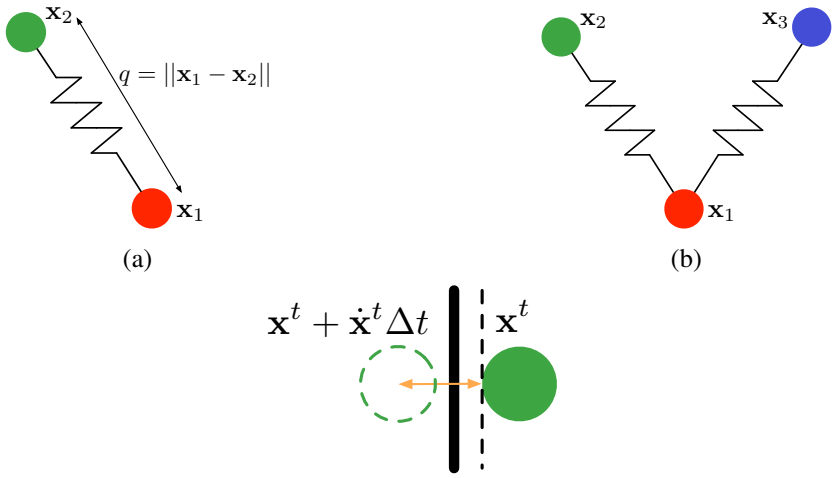


Figure 2: Systems with circles and springs. (a) Two entities (circles) connected by a massless spring. The Cartesian coordinates of the two entities are $\mathbf{x}_1$ and $\mathbf{x}_2$. The potential energy of this system can be defined using just one variable, i.e., the distance between the two entities. (b) Three entities connected by two massless springs. (c) A circle bouncing off a wall. The generalized coordinate in this case can be derived as the expected violation after a short period of time Δt based on the entity's current position $\mathbf{x}^t$ and velocity $\dot{\mathbf{x}}^t$.

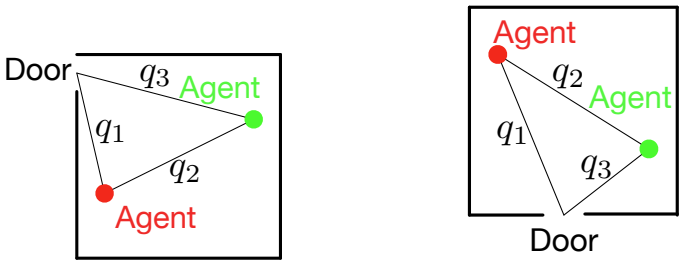
each potential function as $U_j(q_j) = \mathbf{w}_j^\top [q_j, q_j^2]$, where $\mathbf{w}_j$ are polynomial coefficients that can be learned from training examples.

Note that we can define other forms of potential energy functions to capture momentary forces with limited effects depending on spatial ranges. An example of such a case is the momentary forces involved in collision, e.g., an object receives forces momentarily when bouncing off a wall as shown in Figure 2c. To capture such dynamic situations, we relaxed the non-overlapping constraint for rigid objects to allow small spatial protrusion of the entities (the distance between the object and the wall can not be smaller than a small threshold) in a very short period of time (Δt) based on its current position and velocity. Within this spatial range, there will be an effective potential function applied to the entity. In fact, this potential function can be approximated by a spring connecting the contact point of the wall and the object with a very large spring constant $k \gg 1$ and a equilibrium length of distance threshold. This type of approximation has been previously introduced in robotics (Farnioli et al., 2015). If we denote $\delta_j(q_j)$ as the triggering condition function to model the spatial threshold, the complete potential energy can be defined as

$$U(\mathbf{q}) = \sum_{j=1}^{D} \delta_j(q_j) U_j(q_j). \tag{2}$$

2.2 Goal-oriented potentials as value function

In social events, an agent can exert self-propelled forces to purse its goal. To represent goal-directed motion of agents, we extend the PSF model to incorporate goals as a constraint to capture the forces that a rational agent would exert to self-propel in pursuit of its goal. Specially, the model represents the agent's plan with respect to a certain goal as a social force exerted by the agent given its current state and context. The concept of social force has been proposed and applied in previous research (Helbing & Molnar, 1995; D. Xie et al., 2017). However, in prior work social forces are usually limited to manually-specified terms. Rather than using social



(a) (b)

Figure 3: Illustration of social variables as generalized coordinates. (a) An example of generalized coordinates for representing social behaviors in a situation in which one agent aims to leave the room and a second agent attempts to block the first from leaving the room. The (q_1, q_2, q_3) here are the most critical variables in describing this social system. q_1 and q_3 are guided by the potential goal (i.e., the door) for both agents, so an attraction potential term can explain the behavior of "leaving the room." q_2 represents the distance between the agents. For example, a "chasing" relation between two agents could be modeled by a potential term that depends on q_2. (b) Another example showing that when the environment changes (e.g., the door location is altered), we can preserve the same definitions of the generalized coordinates and the potential energy functions, and only modify the transformation from raw observations to generalized coordinates.

forces pre-defined by modelers, the PSF model learns to form Lagrangian representations of social forces from a small set of training data.

The positions of entities, coupled with the goal context, describe the input states of agents. The positions are converted to generalized coordinates to represent socially relevant relations (e.g., the distance between two people) and goal-related variables (e.g., the distance to a door if the goal is to leave a room). An agent's behavior is guided by a potential energy function conditional on a specific goal, which is used to derive the forces governing the dynamics in social events. This goal-conditioned potential function $U_g(\mathbf{q})$ can be viewed as the negative counterpart of a value function, i.e.,

$$V(\mathbf{x}_i|X_{-i}, g) = V(\mathbf{q}, g) = -U_g(\mathbf{q}), \tag{3}$$

where X_{-i} represents states of agents excluding agent i itself. The value function reveals an agent's optimal plan for a goal g and other agents' states – intuitively, a rational agent would want to move to a state with high value (Baker et al., 2017), which indicates high expected return (accumulated rewards) from that state to the future if the optimal plan is followed. For instance, in Figure 3a, if the red agent tries to leave the room, then its motion will be primarily driven by the value function on the first variable $V(q_1)$. Similarly, if the green agent aims to catch the red agent, then it is driven by the value function on the second variable $V(q_2)$. Conveniently, our definition of value function is in relation to a potential function, so that i) by taking the derivative, we can easily derive an agent's optimal policy as social force, ii) such value functions can be learned from a small set of examples in the exact same way as physical potential functions, and iii) the corresponding generalized coordinates capture social variables relevant to an agent's behavior. As we show in our learned social models (Figure 1 and Figure 9), these social variables can reveal an agent's goal position and/or how its plan should be adapted with respect to another agent's plan. Here, we formally define the value function and social force for a given goal g as

$$\hat{\mathbf{F}}_i = \frac{dV(\mathbf{x}_i|X_{-i}, g)}{d\mathbf{x}_i} = \sum_{j=1}^{D} \frac{\partial V(\mathbf{q}, g)}{\partial q_j} \frac{\partial q_j}{\partial \mathbf{x}_i} \quad \forall i = 1, \cdots, N. \tag{4}$$

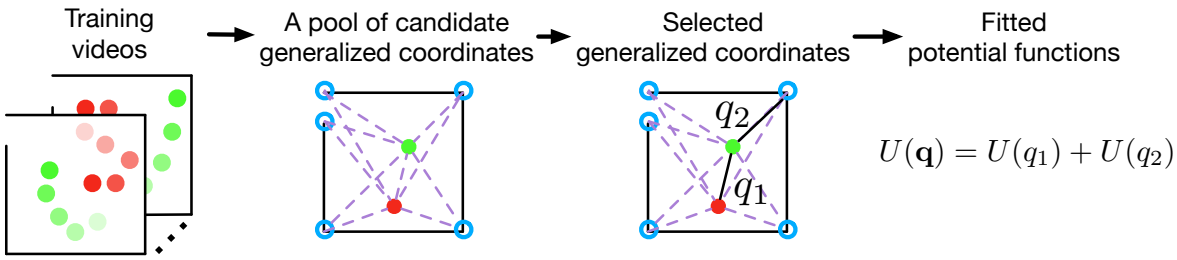


Figure 4: An illustration of the learning components of the PSF model. The model selects a minimum set of generalized coordinates from a pool of candidate coordinates (e.g., the distance between entities and landmarks), and estimates parameters for the potential/value functions. The candidate coordinates are shown as dashed line segments, and the selected coordinates are shown as solid line segments in the figure.

2.3 Learning physical and social models from examples

The input to the PSF model is the motion trajectories of entities (See the stimulus synthesis section for stimulus generation details). A pool of candidates of generalized coordinates, $\mathbb{Q} = \{q_j\}_{j=1}^{D}$, is proposed to the model. The goal of learning is to select a minimal set of generalized coordinates and fit the corresponding potential energy functions, each of which is defined as $U_k(q_k)$, $k \in \mathbb{S} \subset \mathbb{Q}$, where $\mathbb{S}$ is a set of generalized coordinates selected from the candidate pool $\mathbb{Q}$. See appendix for the details about proposals of generalized coordinates. The inferred generalized coordinates and potential functions will then be used to predict the forces for each entity as defined in Eq. 1 from a physical model and Eq. 4 from a social model.

The PSF model learns physical and social knowledge from a small set of training events. For physical events, the PSF learns an ensemble of physical models to capture the characteristics of different physical systems from two types of training examples: 50 videos of collision events, and 50 videos of two objects connected with an invisible spring. For social events, the PSF model learns social knowledge to encapsulate the optimal behaviors of an agent pursuing one of the two specific goals from two types of animations: 50 videos of an agent leaving the room through a door position with a goal of "leaving the room," and 50 videos of an agent attempting to block another agent with a goal of "blocking." For each of the four animation sets, the PSF model selects the appropriate perceptual variables as generalized coordinates and learns potential energy functions that yield the forces that best predict the movements observed in the training data.

Figure 4 illustrates how the PSF model learns generalized coordinates and potential/value functions from a set of training instances (e.g., physical or social events). The learning objective is to discover a minimum set of generalized coordinates and infer the corresponding potential/value functions so that the derived forces can predict the observed movements in the training data. Here, the learning algorithm selects from a fixed set of candidate generalized coordinates (including distance between two entities and distances to the environment landmarks). For the potential/value functions, the learning algorithm estimates the polynomial coefficients for each function to capture the trained physical and social events.

The training data provides the ground truth of forces that each entity receives/exerts at each time point for physical and social events. We define an MSE loss for the force prediction from the PSF model as the learning objective to minimize the discrepancy between predicted and observed forces in the training data:

$$L(\mathbb{S}, \Omega = (w_k)_{k=1}^{K}) = \mathbb{E}\left[\frac{1}{2}||\mathbf{F}_i^t - \hat{\mathbf{F}}_i^t||_2^2\right]. \tag{5}$$

The pursuit of the learning in the PSF model is essentially the search of the optimal generalized coordinates $\mathbb{S}$ and the parameters Ω of the corresponding potential energy functions that

minimize the above loss (along with some regularization for sparsity). For computational efficiency, we adopt a greedy pursuit, which starts from an empty set of generalized coordinates. At each iteration, the learning algorithm augments the candidate generalized coordinate that has not yet been selected in previous iterations and yields a fitted potential energy function to achieve the largest loss reduction from the previous iteration. The iterative pursuit is repeated until there is no significant loss reduction anymore. In the present work, the pursuit is terminated if the loss is no greater than 0.001. Note that given the same training data and the termination condition, the learning results would be unique since the learning algorithm selects the generalized coordinates in a greedy way, which makes the pursuit process deterministic.

In general, we favor sparsity for the coefficients of polynomial terms in the potential or value functions. To learn efficient yet still accurate representations for physical laws, we used Ridge regression. However, to discover parsimonious value functions for social behaviors, we used Lasso regularization instead to pursue a stronger sparsity, which ensures that, for social events, learning sparse value functions with corresponding generalized coordinates provides a coherent framework for explaining the rational behaviors demonstrated by the agents. This approach allows us to infer the optimal policy directly from the learned value function conditional on specific goals. This method also helps us discover sub-goals (i.e., different value function terms) in planning. Finally, the explicit modeling of generalized coordinates can potentially improve the generalization of the learned optimal plans as well, since we can simply remap any new environment to the same generalized coordinate system by only changing the transformation $q(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ while preserving the previously learned value functions. For instance, the generalized coordinates and potential energy functions constructed based on the environment in Figure 3a can be transferred to the new scenario in Figure 3b where the new position of the door will only affect the coordinate transformation for q_1 and q_3. From empirical testing results, we found that the learning in the PSF model was efficient and robust. The algorithm reached convergence within 10 iterations, selecting a small set of physical and social variables that can best represent the generalized coordinates with high sparsity.

2.4 Inference

Based on the generalized coordinates and potential energy functions acquired from learning, the PSF can predict the velocity of entities for new animations using learned physical and goal-directed social models. Learned physical models thus provide a basis for quantifying the discrepancy between observed movements and physical predictions, while learned social models quantify consistency between observed movements and goal-directed predictions. We define the index of *deviation from predictions of physical model* as the mean squared error between observed velocity and predicted velocity from the learned physical model over time. As the physical models are trained with two types of physical events (collision and spring), the PSF uses the physical model that provides the smallest discrepancy between observed and predicted motion to compute the index of deviation from predictions of physical model.

For social events that are governed by two possible goals (leaving the room and blocking the other entity), The PSF model can predict the force fields that would yield the expected motion directions at each location given a specific goal of the agent and the position of the other agent. The predicted fields derived from the goal-directed model provide explicit representations of social forces. Specifically, we compute the log-likelihood ratio of a trajectory following the predicted social forces for pursuing a goal relative to a trajectory of random motion in the absence of goal-directed planning. By choosing the highest log-likelihood ratio among two possible goals (i.e., inferring the most likely goal), we define an index of *degree of goal-directed intention* to provide an intuitive assessment of the impression of intention. See the appendix about the details about mathematical definition of the two indices.

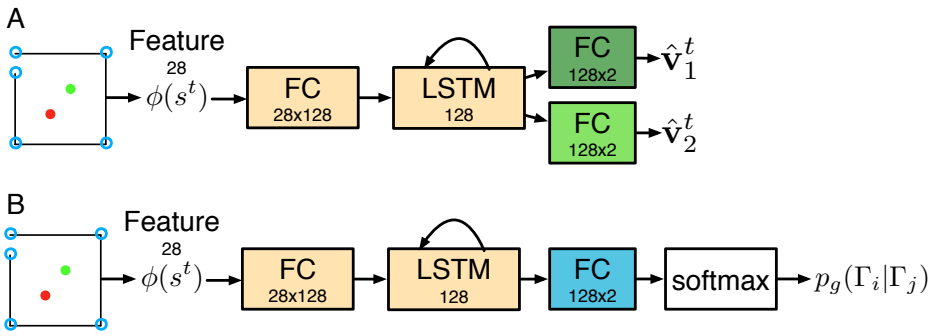


Figure 5: (A) Network for the physical motion prediction model to emulate intuitive physics. Blue circles indicate the corners of the room used for deriving the input features. (B) Network for inferring whether an agent is pursing a specific goal (one network for each goal).

2.5 Baseline model for model comparison: deep neural network

For the purpose of model comparison, we implemented a baseline model by using two deep neural networks (DNN) to measure the deviation from predictions of physical model and the degree of goal-directed intention of an entity in each animation. The DNN baseline model was designed to share the same architecture as the deep reinforcement learning algorithm that was used to generate social events, except using a different output layer. This design was to provide the best opportunity for the DNN model to succeed with the task, as the same architecture is likely to learn representations to capture the patterns of the generated stimuli. To best compare the baseline DNN model and the PSF model, both models shared the same training data and testing inputs. Hence, the performance difference in the two models reflects the representational power learned with the unified framework in the PSF model in comparison to the representation learning in the DNN baseline model. Different from the PSF model, this baseline model learns force representations implicitly through end-to-end training. Each DNN employs an architecture with a long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) to encode the trajectories of entities. Figure 5A shows the network architecture for the physical DNN that learns to predict motion trajectory from the same training examples of physical events used for training the PSF model. Using the DNN predicted motion trajectories, we can thus compute the index of physical violation. For the degree of goal-directed intention, we trained a social DNN with a similar architecture (Figure 5B) for each goal to predict the probability that an agent is pursing that goal, $p_g(\Gamma_i|\Gamma_j)$. We used the same social event training data as for the training of the PSF model, where the trajectory of the agent that is pursing the goal is labeled as a positive example and the other agent's trajectory is labeled as a negative example. This probability is then used to approximate the log-likelihood ratio in Eq. 10, i.e., $\log p_g(\Gamma_i|\Gamma_j) \approx \log p(\Gamma_i|\Gamma_j, g) - \log p(\Gamma_i)$. This method enables us to compute the index of intention. Note that one may also train a single network for goal prediction, which outputs the probabilities for all goals. Our particular design is to match as closely as possible the procedure by which our PSF model computes the index of intention.

3 Stimulus synthesis for model evaluation and human experiments

The animation videos used in previous research usually were generated by manually designing motion trajectories of shapes, or dichotomy rules (e.g., inclusion of self-propulsion movements). Although the films can eliminate confounding factors to distill the precise rules that humans appear to use for interpreting physical and social events, it is hard to design hundreds of different animations to reflect the rich behavioral patterns resulting from the physical and social structure of the world. To train and systematically evaluate the PSF model and the control DNN model, we

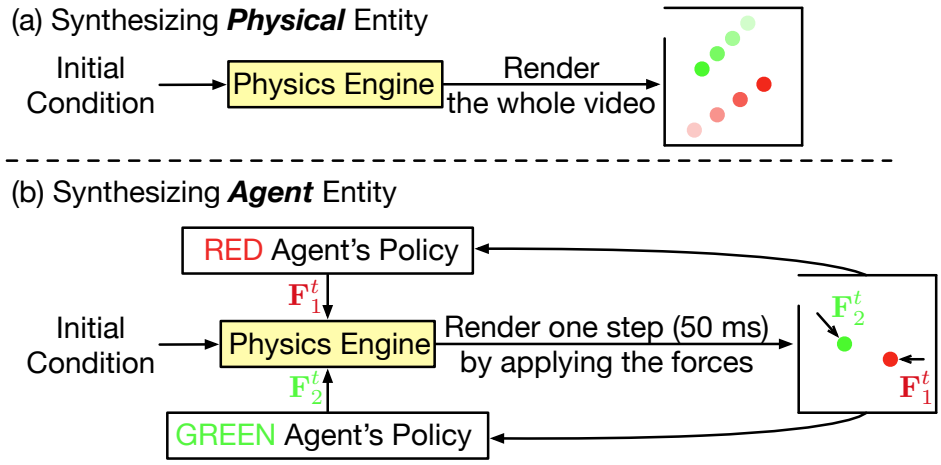


Figure 6: Overview of our joint physical-social simulation engine. For a dot instantiating an inanimate object, we randomly assign its initial position and velocity and then use the physics engine to simulate its movements. For a dot instantiating a human agent, we use policies learned by deep reinforcement learning to provide the forces as inputs to the physics simulation engine.

need a large number of Heider-Simmel-type animations in which movements of simple shapes vary in degrees of deviation from physical predictions and the involvement of goal-directed intention. Therefore, we developed a joint physical-social simulation-based approach by integrating a two-dimensional physics engine and deep reinforcement learning. We used the same simulation engine to generate training stimuli for model simulations and testing stimuli for human experiments.

3.1 Joint physical-social simulation engine

Figure 6 shows an overview of the joint physical-social simulation engine. Each video included two dots (red and green) and a box with a small gap indicating a room with a door. The movements of the two dots were rendered by a 2D physics engine (pybox2d*). If a dot represents an object, we randomly assigned its initial position and velocity, and then used the physics engine to synthesize its motion. Note that our simulation incorporated the environmental constraints (e.g., a dot can bounce off the wall, the edge of the box) and hidden physical constraints between entities (e.g., a rigid rod connecting the dots), but did not include friction. When a dot represented an agent, it was assigned with a clearly-defined goal (either leaving room or blocking the other entity from leaving the room) and pursued its goal by exerting self-propelled forces (e.g., pushing itself towards the door). The self-propelled forces were sampled from agent policy acquired by deep reinforcement learning model (see more details in the next subsection). Specifically, at each step (every 50 ms), the agent observed the current state rendered by the physics engine, and its policy determined the self-propelled force required to advance the agent's pursuit of its goal. We then programmed the physics engine to apply this force to the dot, and rendered its motion for another step. This process was repeated until the entire video was generated.

The ground-truth forces in the stimuli are determined jointly by the external forces (caused by collision and/or invisible physical constraints connecting the entities) simulated by the physics engine, and by self-propelled forces sampled from a trained RL policy for agents. The combined ground-truth forces can be read out from physics engine conveniently and serve for model training to minimize the discrepancy between predicted forces and ground-truth forces in the stimuli.

*https://github.com/pybox2d/pybox2d

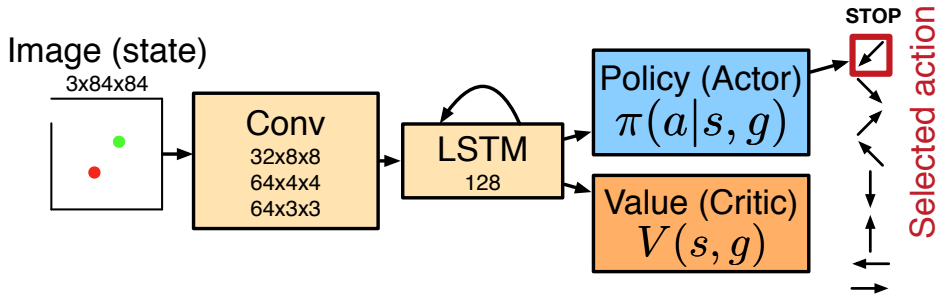


Figure 7: The deep RL network architecture for learning policy for goal-directed movements of an agent. For each goal, we train a separate network with the same architecture.

3.2 Deep reinforcement learning for generating social events

In order to generate social events, we need sensible policies to plan the self-propelled forces for pursuing goals. However, searching for such policies in a physics engine is extremely difficult. We use deep reinforcement learning (RL) to acquire such policies. Deep RL has been shown to be a powerful tool for learning complex policies in recent studies (Silver et al., 2017). Formally, an agent's behavior is defined by an Markov decision process (MDP), $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, \mathcal{G}, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state space (raw pixels as in Figure 7) and action space, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$ are the transition probabilities of the environment (in our case, deterministic transitions defined by physics), R is the reward function associated with the intended goals $g \in \mathcal{G}$, and $0 < \gamma \leq 1$ is a discount factor. To match to the experimental setup, we define two reward functions for the two goals: i) for "leaving the room," the agent receives a reward, $r^t = R(s^t, g_1) = \mathbb{1}(\text{out of the room})$, at step t; ii) for "blocking," the reward at step t is $r^t = R(s^t, g_2) = -\mathbb{1}(\text{opponent is out of the room})$. To simplify the policy learning, we define a discrete action space, which corresponds to applying forces with the same magnitude in one of the eight directions and "stop" (the agent's speed decreases to zero after applying necessary force).

The objective of the deep RL is to train the policy of the system (see its architecture in Figure 7) to maximize the expected return $E[\sum_{t=0}^{\infty} \gamma^t r^t]$ for each agent. Thus, for "leaving the room," the learnt policy would enable the agent to quickly move toward the door. For "blocking," the best policy relies on the position of the second agent and would enable successful blocking. The optimization is implemented using the advantage actor critic (A2C) (Mnih et al., 2016) method to jointly learn a policy (actor) $\pi : \mathcal{S} \times \mathcal{G} \mapsto \mathcal{A}$ that maps an agent's state and goal to its action, and a value function (critic) $V : \mathcal{S} \mapsto \mathbb{R}$.

3.3 Generation of training stimuli

We used the following parameter settings to generate the training stimuli. Social events in training include 50 videos of an agent leaving the room through a door position, and 50 videos of an agent attempting to block another agent with a goal of "blocking." The parameters of agent policies were acquired by deep reinforcement learning described in the above section to sample self-propelled forces governed by the agent's goals. For physical events (50 videos of collision events, and 50 videos of two objects connected with an invisible spring), dots represent objects without goals. The movements of dots follow physical rules. For collision events, the initial positions and velocities of dots were randomly assigned; for movements of objects connected with an invisible spring, the frequency was set to be 0.3. To generate the training stimuli, the forces (i.e., the combined self-propelled force sampled from goal-directed policy and the external force) were applied to the dots in the physics engine to generate the movements of dots. We have released the stimulus generation code and the trained model for generating all stimuli at `https://github.com/MicroSTM/HeiderSimmelRL`.

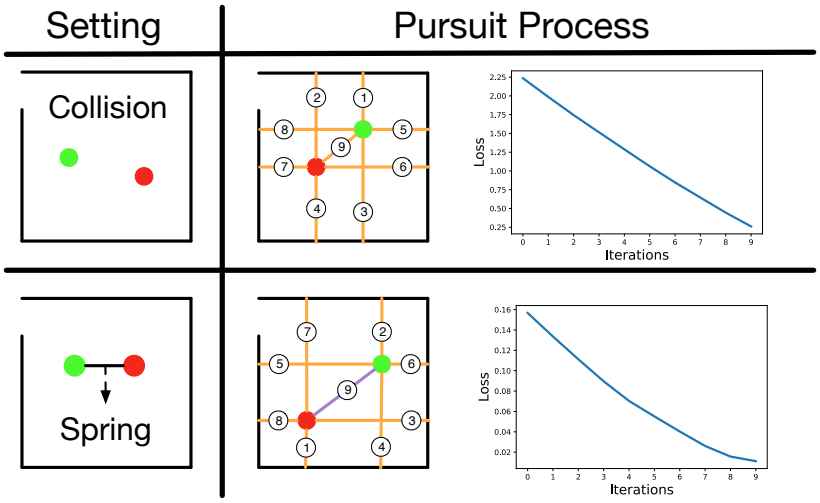


Figure 8: Illustration of learning results of two physical systems. The purple and orange lines in the middle plots are the selected generalized coordinates from the first and the second type of candidates, respectively; each number indicates the iteration step when the corresponding generalized coordinate was selected. The right plots show that adding more generalized coordinates reduces the MSE loss between the predicted forces and the ground-truth forces.

4 PSF learning results and simulation results for classic stimuli

First, we show learning results of the PSF model from some example animations. Figure 8 depicts the generalized coordinates and the learning pursuit process for two physical systems: collision and spring (with several different spring lengths), each trained with 50 examples of animations. For events with social goals, the same learning approach was implemented to pursue value functions for two goals (i.e., leaving the room and blocking). Figure 9 shows generalized coordinates and the derived forces fields based on the final step of the learning for each goal.

To test whether the PSF model can be applied to untrained stimulus types, we applied the learned PSF model to five classic animations used to study causal perception, adopted from (Scholl & Tremoulet, 2000). Figure 10 shows the trajectory of motion inputs from the stimuli of "entraining," "triggering," "launching," and "launching" with spatial and temporal gaps. Although the stimulus changes are subtle in these animations, human perception changes dramatically. The "launching" stimulus generates irresistible impression of a physical event that one ball causally launched the other ball (Michotte, 1963; Scholl & Tremoulet, 2000). But when a small spatial or temporal gap is added to the launching stimulus, the impression of a collision event attenuated significantly (Michotte, 1963). Interestingly, the "triggering" and "entraining" animations appear to yield an impression consistently reported by human observers as intentionally moving away and pushing a target (Hubbard, 2013; Scholl & Tremoulet, 2000). These perception changes are well captured by the PSF model. The motion trajectories in each stimulus were input to the PSF model trained with physical events and social events described in the 2.3 section. The PSF model then derived the two indices in the sociophysical space for each stimulus. The right panel in Figure 10 depicts the placement of each stimulus in the model-derived sociophysical space. The "entraining" stimulus shows the highest degree of goal-directed behavior because the entrainer (red ball) continues to move after contacting the target green ball, and the two objects remain in contact to yield impression of pushing action. The "triggering" stimulus shows the largest deviation from physical predictions, as the green ball speeds up after the collision contact which is in contradictory to physical predictions. The "launching" stimulus (standard version)

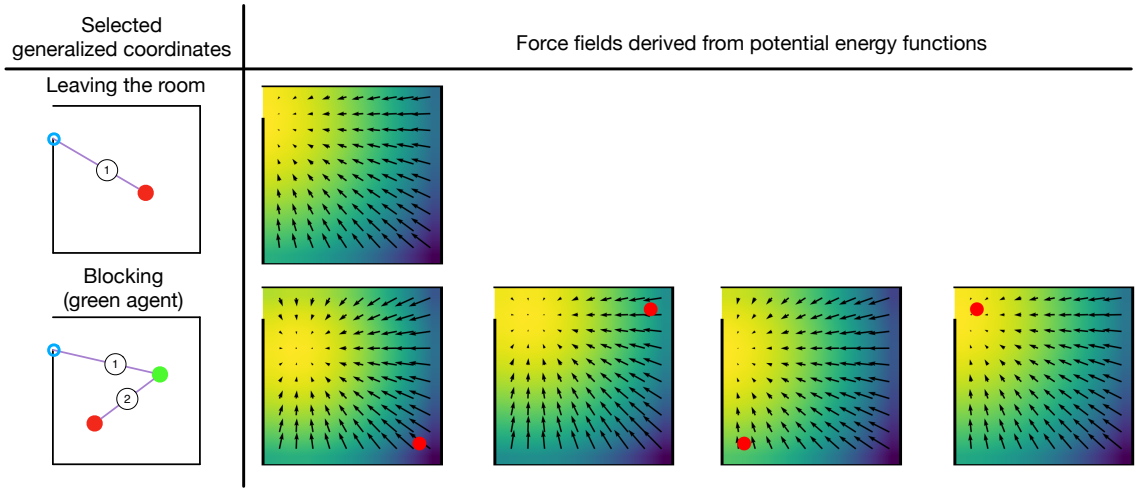


Figure 9: Learning results of two social events each with different goals. Left: selected generalized coordinates; right: the learned value functions and force fields derived from the value functions, where the red circle represents the position of the other agent, and the color of the background indicate the value of a state (blue to yellow indicates low value to high value). An agent will move towards high value positions and move away from low value positions.

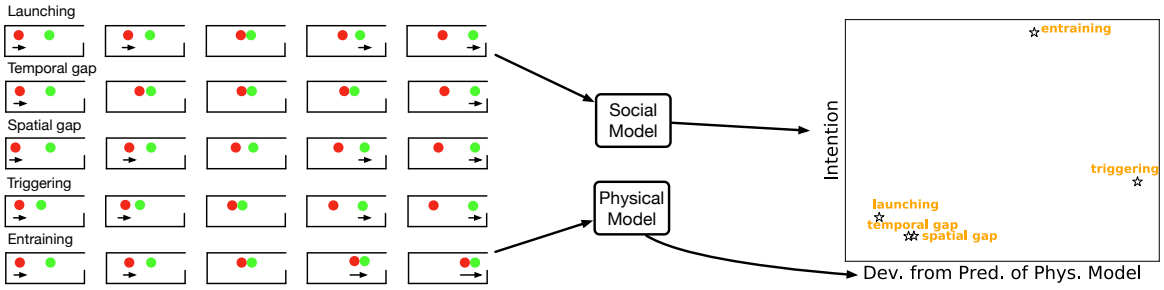


Figure 10: Simulation results for five classic stimuli used to study causal perception research (Scholl & Tremoulet, 2000), showing the resulting location of each stimulus in the unified space.

shows high consistency with physical predictions to be located in the bottom left corner of the space, and the "launching" stimulus with spatial or temporal gaps exhibit some deviations from physical predictions. These intuitively reasonable results of the application of the PSF model to classic stimuli from the psychology literature provide preliminary support for the computational framework. Next, we report two psychophysical experiments designed to quantitatively test whether the PSF model can account for human judgments across a large range of stimulus conditions.

5 Experiment 1: Identification of interaction types

In Experiment 1, we generated hundreds of animations depicting a wide range of behaviors representative of different physical and social scenarios. Human performance on the identification of interactions types were collected and were used to compare with PSF model predictions.

5.1 Methods

5.1.1 Participants

Thirty University of California, Los Angeles (UCLA) undergraduate students (mean age = 20.9; 19 female) participated in the experiment for course credit. All participants had normal or corrected-to-normal vision. Participants were provided written consent via a preliminary online survey in accordance with the UCLA Institutional Review Board and were compensated with course credit.

5.1.2 Stimuli

Figure 1 summarizes the three major types of interactions that were examined: object-object (OO) , human-object (HO), and human-human (HH) interactions, all of which were generated by the joint physical-social simulation engine.[†] The environment consists of a square box with a gap indicating the exit corner. Two dots (one red and one green) move in the box, and only can move out of the box from the exit corner. To synthesize motion of dots as agents, we set two types of goals for the agents: "leave the room" and "block the other entity." Specially, in HH animations, one agent has a goal of leaving the room, and the other agent aims to block it; in HO animations, an agent with a goal of "blocking" always attempts to keep a moving object (an initial velocity towards the exit corner) within the room. By randomly assigning initial position/velocity to the two dots and stochastic control through reinforcement learning, we can simulate rich behaviors that give subjective impressions such as blocking, chasing, attacking, and pushing. Videos lasted from 1 s to 1.5 s with a frame rate of 20 fps. We synthesized a total of 850 videos of Heider-Simmel-type animations without filtering, with 500 HH videos, 150 HO videos, and 200 OO videos.

For HH videos, we introduced different degrees of animacy by varying how often a dot seemed to exert self-propelled forces in an animation. In general, a higher degree of animacy is associated with more frequent exertion of self-propelled forces directed towards the goal, and animations with a low degree of animacy are more likely to be perceived as physical events. This manipulation generated five subcategories of HH stimuli with five degrees of animacy: 7%, 10%, 20%, 50%, and 100%, respectively corresponding to applying force once every 750, 500, 250, 100, or 50 ms. In an HH animation, we assigned the same degree of animacy to both dots. There were 100 videos for each level of animacy degree for the HH animations. It should be noted that in training, the model was only presented with videos with the full degree of animacy and never saw other conditions.

In OO animations, we included four physical events as shown in Figure 1: a collision between two rigid objects, two objects connected with an invisible rod, with an invisible spring, or with an invisible soft rope, with 50 videos for each sub-category. Because these connections were invisible in the display, the hidden physical forces might result in a subjective impression of animacy or of social interactions between the entities. In addition, the invisible connections between objects (rod, spring, and soft rope) introduced different degrees of violation to the movements expected under the assumption of simple physical interactions (e.g., the two physical entities are not connected and hence move independently). We controlled the initial velocities to ensure that the average speeds of dots in OO animation videos were the same as the average speeds of dots in HH animation videos (44 pixel/s).

5.1.3 Procedure

The dataset was split into two equal sets; each contained 250 HH, 75 HO, and 100 OO videos. 15 participants were presented with set 1 and the other 15 participants were presented with set 2. Stimuli were presented on a 1024×768 monitor with a 60 Hz refresh rate. Participants were given the following instructions: "In the current experiment, imagine that you are working for a

[†]Some example videos can be viewed at `https://youtu.be/3KmhjB-chvM`.

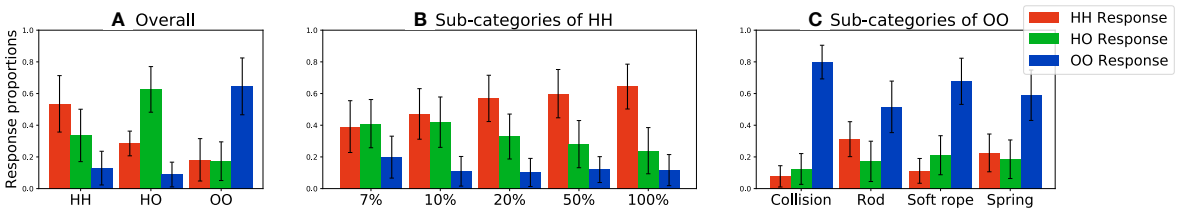


Figure 11: Response proportions for human judgments in Experiment 1. (A) major interaction categories; (B) subcategories of HH videos differing in degree of animacy; (C) subcategories of OO videos differing in type of physical interaction. Error bars indicate standard deviations across stimuli.

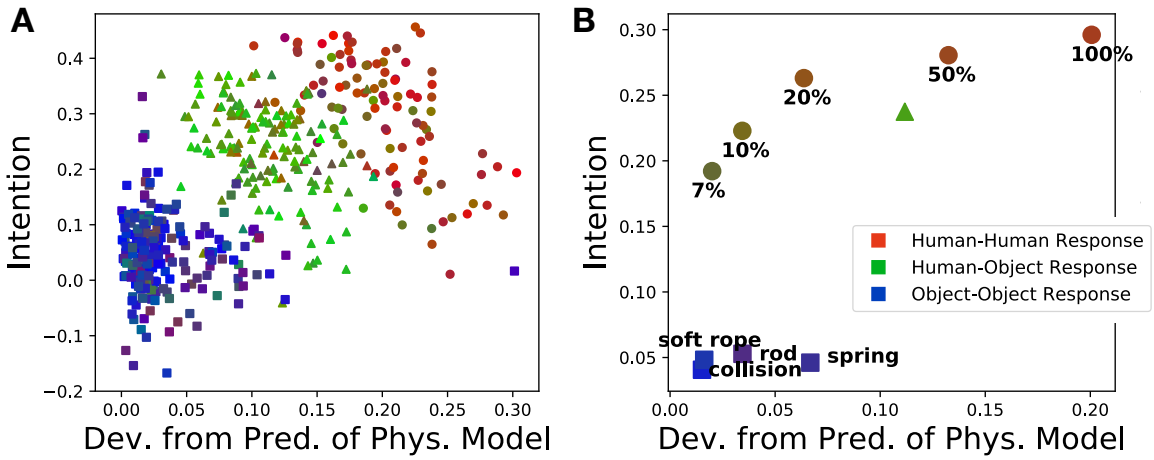


Figure 12: Results of Experiment 1 in the sociophysics space. (A) Results of PSF model showing constructed psychological space for HH animations with 100% animacy degree, HO animations, and OO animations. A stimulus is depicted by a data point with coordinates derived by the model. The mark shapes of data points correspond to the interaction types used in the simulation for generating the corresponding stimuli (circle: HH, triangle: HO, square: OO). The colors of data points indicate the average human responses for this stimulus. Reddish color indicates more human-human responses; greenish color indicates more human-object responses; and bluish color indicates more object-object responses. (B) Average coordinates of indices predicted by the PSF model for videos in each of the sub-categories in the constructed psychological space, including HH animations (circles) with different levels of animacy degree, and OO animations (squares) with four subtypes of physical events. The colors of center points indicate the average human responses of stimuli in this subcategory. Reddish color indicates more human-human responses; greenish color indicates more human-object responses; and bluish color indicates more object-object responses.

security company. Videos were recorded by bird's-eye view surveillance cameras. In each video, you will see two dots moving around, one in red and one in green. There are three possible scenarios: human-human interaction, human-object interaction, or object-object interaction.Please pay attention to movements of the two dots and judge what kind of interaction the two dots demonstrate after each video. There is no right or wrong regarding your decisions. Press Left-Arrow button for Human-human, Up-Arrow button for Human-object, and Right-Arrow button for Object-object." Videos were presented in random orders. After the display of each video, participants were asked to classify the video into one of the three categories. The experiment lasted for about 30 minutes.

5.2 Results and discussions

Human response proportions are summarized in Figure 11A. For each stimulus type, the majority of responses correctly identified the category of the interaction depicted in the videos produced by the simulation engine (49% responses of human-human interaction for HH animations, 62% responses of human-object interaction for HO animations, and 68% responses of object-object interaction for OO animations). However, human judgements also revealed considerable uncertainty. On average, about 40% responses were interaction types other than that used for stimulus generation. For example, for HH animations 39% of responses selected human-object interactions and 12% responses selected object-object physical interactions. For OO animations, 16% of participants considered these animations to be human-human social interactions.

Figure 11B depicts the impact of degree of animacy on judgments for HH animations. Degree of animacy was manipulated by varying the frequency with which an agent exerted self-propelled force directed towards the goal. With increasing degrees of animacy, the proportion of human-human interaction responses also increased, yielding a positive Pearson correlation with the frequency of applying self-propelled forces towards a goal ($r = .42$, $p < .001$). When the degree of animacy was low (only 7% frequency of goal-directed forces added to the movements), people were more likely to classify the animations as human-object interactions or physical interactions.

As shown in Figure 11C, the response proportions suggest uncertainty in human's judgments on identifying physical interactions across the four subcategories of OO animations. A one-way ANOVA showed that the proportion of object-object interaction responses differed significantly among the four subcategories of physical interactions ($F(3,196) = 34.42$, $p < .001$, $\eta^2 = .345$), with the most object-object responses in the collision condition, and the least in the connected-with-invisible-rod condition. Interestingly, the proportion of human-human interaction responses for OO animations also yielded significant differences among the four subcategories ($F(3,196) = 61.15$, $p < .001$, $\eta^2 = .483$). Specifically, invisible-rod and invisible-spring conditions yielded a higher proportion of human-human interaction responses than did collision conditions, suggesting that some invisible physical constraints between objects may give rise to a mistaken impression of human-human social interactions.

For each video shown to human observers, the PSF model computes the two indices to measure how well the observed motion patterns can be predicted by the physical model (the index of deviation from predictions of physical model), and the likelihood that dots are agents showing intentions (the index of degree of goal-direction intention). If these two dimensions capture the key computational components underlying human intuitive impressions about physical and social events, then the theoretical space should show clear partitions of different types of animations that are perceived by human participants.

Figure 12A depicts the PSF model-derived results and human responses for 100 HH videos with 100% animacy degree, 150 HO videos, and 200 OO videos. The position of each video in the space was determined by the estimated indices assessing the discrepancy from physical predictions (horizontal) and the degree of consistency with intention-based predictions (vertical). The coordinates of each data point were calculated as the model-derived measures averaged across the two entities in an animation. The mark shapes of data points correspond to the interaction type that was used for generating stimuli by the joint physical-social simulation engine. The colors of data points indicate average human responses for the stimulus. Specifically, the values of RGB channels are determined by the average human-human responses in red, human-object responses in green, and object-object responses in blue. The resulting space shows clear separations between the animations that were judged to be each of the three different types of interactions. Animations with more human-human interaction responses (reddish marks) clustered at the top-right corner, corresponding to high values of intention index and strong evidence signaling violation of physical model. Animations with high responses for object-object interactions (bluish marks), located at the bottom left of the space, show low values of the intention index and little evidence to deviation from predictions of physical model. Animations with high responses for human-object interactions (greenish marks) fell in the middle of the space.

To quantitatively evaluate how well the two indices in the unified sociophysics space account

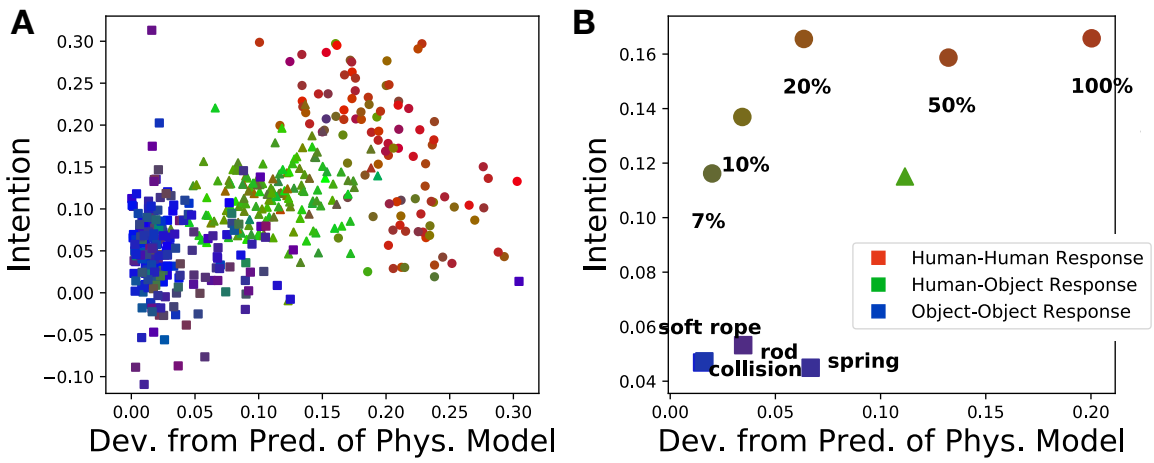


Figure 13: Results of Experiment 1 in the sociophysics space based on the ablated model which used all candidate coordinates. (A) The distribution of individual animations in the 2D space based on the physics and intention inference by the ablated model. (B) Centers of the sub-categories in the 2D based on inference results by the ablated model.

for human judgments, we trained a cluster-based classifier using the coordinates of each animation in Figure 12A as input features. The correlation between the model predictions and average human responses across the trials was 0.787 based on 2-fold cross-validation. Using a split-half reliability method, human participants showed an inter-subject correlation of 0.728. Hence, the response correlation between model and humans closely matched inter-subject correlations, suggesting the unified sociophysics space provides a general account of human perception of physical and social events based on movements of simple shapes.

To evaluate the benefit of pursuing a parsimonious model, we conducted an ablation study by training the PSF model using all candidate coordinates. As shown in Figure 13, the resulting model predictions produced a similar but less clear separation in the 2D space, and a slightly lower correlation with human responses (0.747).

We tested the DNN-based baseline model on the same stimuli to construct a two-dimensional space. We used the same clustering methods to predict interaction judgments for animations used in the experiment. The resulting DNN predictions yielded a poor correlation of 0.344 with human responses. The superiority of the force-based PSF model relative to the data-driven DNN models as an account of human judgments suggests that human perception of physical and social events infers hidden forces governing the dynamics of physical and social scenarios, rather than simply using pattern recognition based on previous experience.

We further examined the impact of different degrees of animacy and of different subcategories of physical events on model predictions. The unified space derived by the PSF model provides a platform to compare these fine-grained judgments. Figure 12B shows the center coordinates as the averaged indices for videos in each of the subcategories. Simulation results show that with decreased degree of animacy, both the intention index for HH animations and the index of deviation from predictions of physical model were gradually reduced. Similarly, human judgments for these stimuli varying from high to low degree of animacy transited gradually from human-human responses to human-object responses, consistent with the manner in which the data points moved towards the bottom left corner of the space. Among the subcategories of physical events, the invisible-rod and invisible-spring conditions respectively showed the highest intention index and the strongest physical violation, consistent with the greater proportion of human-human interaction responses from participants.

Based on the observed association between the degree of animacy (i.e., the frequency of trajectory change) and human judgments, we further trained a simple baseline model to directly

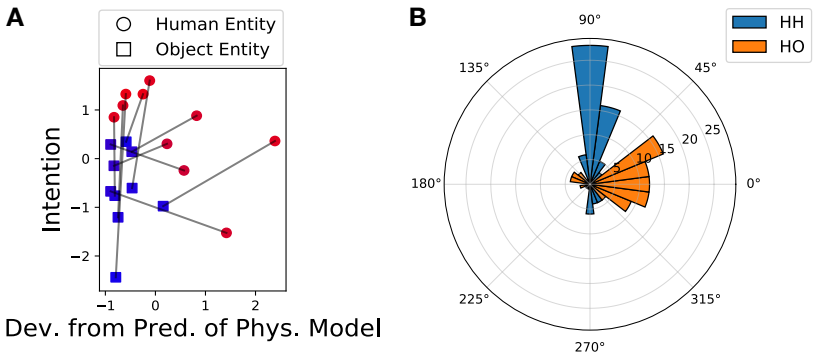


Figure 14: Human and model results in Experiment 2. (A) Representative cases of animations that elicited human-object responses, located in the space with model-derived coordinates. The colors reflect average human responses for assigning a dot to the human role (red) and to the object role (blue). (B) Orientation histogram of the segments connected by the coupled pairs of entities in each animation.

examine whether the degree of animacy alone can predict human responses. Hence, we trained a logistic regression to predict the HH/HO/OO interaction categories based on the frequency of trajectory change for the two dots (i.e., a 2-dim feature vector, each dimension indicating the frequency of a dot). The model predictions from this simple baseline show a very weak correlation ($r = 0.002$) with human responses, suggesting that humans' judgments cannot be explained by the frequency of trajectory change alone. Considering an example where a dot moves towards the door in a straight line, the dot exhibits a strong intention of leaving the room, without demonstrating any change in trajectories. This simple example illustrates the limitation that the frequency of trajectory change is not an effective cue in identifying social events.

6 Experiment 2: Role identification in human-object interactions

If the unified space derived from the PSF model forms efficient representations for both perception of physical and social events, we would expect that the spatial representations in the unified space will predict human judgments in other tasks. In Experiment 2, we examined the task of assigning roles to individual entities. Here, we focused on stimuli that elicited the impression of human-object interaction in Experiment 1, and asked a new group of participants to report which dot was a human agent, and which dot was an inanimate object.

6.1 Methods

Twenty-five new UCLA students (mean age = 20.2; 19 female) were recruited from the UCLA Psychology Department Subject Pool.

The top 80 HH videos and the top 80 HO videos that got the highest response proportions of being judged as human-object interactions in Experiment 1 were used for Experiment 2. The procedure was the same as Experiment 1 except that on each trial, subjects were asked to judge among two dots, which dot represented a human agent and which dot represented an object. One dot was red and the other was green and the colors were randomly assigned to the two dots in each trial. The experiment lasted about 15 minutes.

6.2 Results and discussions

For each animation, the PSF model projected individual entities to the unified sociophysics space based on the indices for each individual entity. Figure 14A shows representative cases of 5 HH animations and 5 HO animations. The coordinates of each entity were derived from the PSF model, and the two entities in each animation were connected to create a line segment. Red circles represent the dots that participants identified as a human agent, and blue squares represent the dots that participants identified as an inanimate object. To make the scale of the two indices directly comparable, we used a standardized score for the indices in Figure 14A.

The resulting segments in Figure 14A revealed a consistent property: the agent dot identified by participants shows a higher degree of goal-directed intention and greater discrepancy from physical predictions than does the object dot identified by participants. This property implies that the orientation from the object entity to the agent entity is within the range of 0 to 90 degree in the unified space. For each animation, we thus used the social and physical indices from the PSF model and the role assignments reported by human participants to compute the orientation from the object entity to the agent entity. The histogram of the calculated orientations is shown in Figure 14B). The orientation histogram reveals a bimodal distribution with one peak around 90 degrees for HH animation videos and another peak between 0 and 45 degrees for HO animation videos. Hence, the majority of the videos showed the role orientation within the 90 degree range, suggesting that role identification for individual entities can be assessed by how well the motion of the entity conforms with physical constraints and social goals that captured in the PSF model. Specifically, for HH animations, movements of both dots show similar discrepancy from predictions of physical models, which makes the index of the deviation from predictions of physical model uninformative for the role assignment. Hence, participants tend to assign the dot with a lower degree of intention as the object role, and the dot with a higher degree of intention as the agent role. Such consistent strategy leads to the clustering of role assignments for HH animations around 90 degrees in Figure 14B). On the contrary, the role assignment for HO animations relies on differences from the degree of intention and the deviation from predictions of physical model. Participants tend to assign the agent role to a dot showing more intention and stronger deviation from predictions of physical model and to assign the object role to a dot revealing less intention and more consistency with predictions of physical models, resulting a second cluster around 45 degrees in Figure 14B).

7 General discussions

The present paper provides evidence that the perception of physical events and social events can be integrated within a unified sociophysics space, and can be modeled using a continuous spectrum to capture both physical and social domains. This common representation enables the development of a comprehensive computational model of how humans perceive physical and social scenes. We showed that a classification based on just two model-derived measures (reflecting violation of physical knowledge and an impression of goal-directed actions) can predict human judgment well, reaching the same level as inter-subject reliability. This good fit to human responses across a range of Heider-Simmel-type stimuli demonstrates the effectiveness of using a unified modeling approach to study the transition from intuitive physics to social perception.

The main benefit of constructing this sociophysics space is to provide an intuitive assessment of general impressions of physical and social events from a brief presentation of sparse visual information. To build up such representations, humans or computation models may use various cues to detect intentions and/or physical violations. The proposed space provides an abstract framework for gauging how humans' intuitive senses of physics and intentions interact to guide perception of physical and social events.

The PSF model derives the unified sociaphysics space by constructing generalized coordinates and the corresponding potential/value functions. There are three advantages of using this representation system to model both physical and social events. First, Lagrangian representations is effective for inferring forces across both physical and social domains. The selection of the

generalized coordinates reveals the effective change of a dynamic system regardless of the source of the forces. By pursing the generalized coordinates that results in the simplest potential and value functions, we are essentially inferring a sparse model for the dynamic system. For physical systems, such representations will reveal physical concepts, whereas in social systems, they reveal goals and social relations. Second, the PSF model enables the "compression" of optimal planning in social events. Optimal planning is complex and time consuming. However, given observed trajectories of agents, the PSF model compresses these optimal plans into a small number of value functions. Consequently, instead of searching for an optimal plan from scratch over time for each animation, we can derive forces from the value functions and roll out the whole plan step-by-step starting from the initial state. We may deploy this plan directly, or use it as a starting point for further refinement to compensate for the deviations between the behaviors predicted from the learned value functions and observed behaviors. Similarly, we can also take advantage of the derived forces to guide inverse planning for Bayesian goal inference, which can significantly reduce computational demands. Third, the PSF model demonstrates generalization ability and makes it possible to transfer learned knowledge to new environments. Even though in the current setting, only two social goals and two physical interaction types were included in the training, the learned force representation and the unified space can adapt to novel scenarios. Even when the surrounding environment changes, the potential energy defined on generalized coordinates can be preserved. In order to derive forces for the entities in the new environment, we only need to change the coordinate transformations. This was demonstrated by the simulation results for the five classic stimuli commonly used in psychological research to study causal perception (Scholl & Tremoulet, 2000), obtained even though the PSF model was not trained with this type of stimuli (involving only horizontal movements in a 1D space).

This work provides an important step toward developing a computational theory to unify human perception and reasoning for both physical and social environments. The physical and social models learned from synthesized animations can be generalized to other stimuli used in previous psychology studies, suggesting a promising scope of generalization. However, the PSF model has limitations. For example, the current simulations include only a small set of goals, and the model requires a pool of predefined goals and accurate knowledge about the physical environment. This limits the generalization of the learned PSF model to a small set of social situations with goals similar to the training events. Future work should aim to extend the analysis to a broader range of goals in social events, to develop more sophisticated modules for goal inference, and to support causal perception in human actions (Peng et al., 2017). The current model also only targeted simplified decontextualized scenarios. A more complete model would possess the ability to learn about physical environments based on partial knowledge, and to emulate a theory of mind in order to cope with hierarchical structures in the goal space. In addition, the current study has only examined human perception of physical and social events for short-duration stimuli involving two entities. Generating videos of longer events with more entities and temporal segments, and analyzing human perception of them, will allow further investigations of the mechanisms underlying physical and social perception in humans.

8 Appendix

8.1 Learning algorithm

Learning input. In a dynamic system with N-entity, we observe the trajectories of all entities $\Gamma_i = \{(\mathbf{x}_i^t, \dot{\mathbf{x}}_i^t)\}_{t=1}^T$, where the length of each time step is Δt, and the total duration is $T\Delta t$. We assume that all entities have the same mass m and only conservative forces are present in the system. From the trajectories, we can compute the ground-truth force each agent i receives at time step t, i.e., $\mathbf{F}_i^t$. The goal is to learn a model (generalized coordinates and potential energy functions) that can predict the forces for the observations.

Proposals of generalized coordinates. We obtain a pool of candidates for generalized coordinates, $\mathbb{Q} = \{q_j\}_{j=1}^D$. Note that many of them may be redundant and will not be selected by the

final model. In particular, these candidates can arise from two types of proposals:

i) Distances between entities involved in animations and environment landmarks. These can be the distance between two entities (e.g., the one in Figure 2) or the distance between an entity and a part of the environment (e.g., the distance between the corners of the room and the dots). The corresponding potential energy functions are always triggered, i.e., $\delta_j(q_j) = 1$.

ii) Expected constraint violation as illustrated in Figure 2c. When there is violation, q_j represents the expected overlapped length; otherwise $q_j = 0$. The triggering condition is consequently defined as $\delta_j(q_j) = \mathbb{1}(q_j > 0)$. This means that the corresponding term would only take effect when the generalized coordinate q_j is positive. Intuitively, this models the triggering condition of certain momentary physical forces, such as collision.

Note that for social events, we do not consider the second type of generalized coordinates.

8.2 Physics inference

By giving the positions and velocities of the two entities at time t, i.e., $\mathbf{x}_i^t$, $\dot{\mathbf{x}}_i^t$, $i = 1,2$, we can use the learned variables (i.e., generalized coordinates) and potential functions to predict the physical forces each entity receives at t and consequently their future velocities at $t+1$, $\hat{\dot{\mathbf{x}}}_i^{t+1}$, $i = 1,2$. By comparing with the ground truth $\dot{\mathbf{x}}_i^{t+1}$, we can evaluate the degree to which an entity's motion is inconsistent with predictions of physical models:

$$\mathcal{D}_i = \frac{1}{T}\sum_{t=1}^{T} ||\dot{\mathbf{x}}_i^t - \hat{\dot{\mathbf{x}}}_i^t||_2^2, \quad \forall i = 1,2. \tag{6}$$

In the PSF model, there are multiple physical situations (e.g., collision or spring), each of which include different sets of generalized coordinates and potential functions to provide various predictions about the movements of entities. Since the PSF model does not know to what system an observation belongs, we enumerate all learned physical situations, and apply a winner-take-all strategy to select the one that yields the best prediction revealed by the lowest MSE error. The selected physical model will be used to derive the measure of physical violation.

8.3 Intention inference

The social force fields yield the expected moving direction at each location given the goal of the agent and the position of the other agent. Inspired by the classic FRAME model (Zhu et al., 1998; J. Xie et al., 2015) which was originally used for modeling texture and natural images, we can view a force field derived from the learned PSF model as motion filters with values changing at different locations given a specific goal. Specifically, the filter response at location $\mathbf{x}_i$ for agent i with goal g_i and the other agent being located at $\mathbf{x}_j$ can be defined as

$$h(\dot{\mathbf{x}}_i|\mathbf{x}_i, \mathbf{x}_j, g_i) = cos(\theta) = \frac{\hat{\mathbf{F}}_i(\mathbf{x}_i|\mathbf{x}_j, g_i)^\top \dot{\mathbf{x}}_i}{||\hat{\mathbf{F}}_i(\mathbf{x}_i|\mathbf{x}_j, g_i)|| \cdot ||\dot{\mathbf{x}}_i||}, \tag{7}$$

where θ is the angle between the observed moving direction $\dot{\mathbf{x}}_i$ and the expected moving direction from the predicted forces $\hat{\mathbf{F}}_i$ as shown in Eq. 4. The comparison of the moving directions makes the model inference robust to speed change. By dividing the whole image space into R discrete regions ($R = 4$ in this work), where each region has a location set $\mathbb{X}_r$, we can define the likelihood of observing an agent with a goal having a certain trajectory Γ_i as

$$p(\Gamma_i|g_i, \Gamma_j) = \frac{1}{Z(\Lambda)} \exp\left\{\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{R} \mathbb{1}(\mathbf{x}_i^t \in \mathbb{X}_r)\lambda_r h(\dot{\mathbf{x}}_i^t|\mathbf{x}_i^t, \mathbf{x}_j^t, g_i)\right\} q(\Gamma_i), \tag{8}$$

where $q(\Gamma_i) = \prod_{t=1}^{T} q(\dot{\mathbf{x}}_i^t)$ is a baseline model for moving directions without pursuing a specific goal. We assume a uniform distribution for $q(\dot{\mathbf{x}}_i^t)$ for the baseline. $\Lambda = (\lambda_1, \cdots, \lambda_R)$ is the parameter for the likelihood corresponding to the R regions, and $Z(\Lambda)$ is the normalization term that can be defined as

$$Z(\Lambda) = E_{q(\Gamma)}\left[\exp\left\{\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{R}\mathbb{1}(\mathbf{x}_i^t \in \mathbb{X}_r)\lambda_r h(\dot{\mathbf{x}}_i^t|\mathbf{x}_i^t, \mathbf{x}_j^t, g_i)\right\}\right]. \tag{9}$$

Since we assume a uniform distribution for the baseline model in the absence of goals, it is easy to show that $Z(\Lambda) = 1$. Parameter λ_r in the likelihood can be estimated from the motion filter responses to trajectories in training examples in region r. Finally, we define the index for intention measurement as the log-likelihood ratio of a trajectory following the optimal plan for pursuing *any* goal over the baseline model in the absence of goals:

$$\mathcal{L}_i = \max_{g \in \mathcal{G}} \log p(\Gamma_i|g, \Gamma_j) - \log q(\Gamma_i), \quad \forall i = 1, 2. \tag{10}$$

8.4 Data availability

The data that support the findings of this study are available at `https://osf.io/5sxuw`.

8.5 Code availability

The code for the stimuli generation is available at `https://github.com/MicroSTM/HeiderSimmelRL`, and the code for the PSF model is available at `https://github.com/MicroSTM/PSF`.

9 Acknowledgement

We thank Ciaran Zhou, Claire Locke, Huiwen Duan, Suhwan Choi, and Zhibo Zhang for assistance in data collection. We also thank Jianwen Xie, Ying Nian Wu, and Tao Gao for the helpful discussions. This research was supported by NSF Grant BCS-1655300 to HL, and by DARPA XAI N66001-17-2-4029 and ONR MURI project N00014-16-1-2007 to SZ.

10 Author contributions

T.S. formulated and implemented the models. T.S., Y.P. and H.L. designed experiments. Y.P. ran human experiments. T.S. and Y.P. performed data analyses for the experiments. S.-C.Z. investigated key ideas and provided funding support. T.S., Y.P. and H.L. drafted the manuscript. S.-C.Z. provided important guidance for refining the manuscript.

11 Competing interests

S.-C.Z. has affiliation with DMAI Inc., a startup company dedicated to education. The other authors declare that they have no competing interests. The research presented in this article is entirely funded by an NSF grant and an ONR MURI project, and is conducted at UCLA.

References

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*(4), 1–10.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

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.

Burge, T. (2018). Do infants and nonhuman animals attribute mental states? *Psychological Review*, *125*(3), 409.

Csibra, G., Gergely, G., Bıró, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason'in infancy. *Cognition*, *72*(3), 237–267.

Danks, D. (2009). The psychology of causal perception and reasoning. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The oxford handbook of causation.* Oxford University Press.

Dittrich, W. H., & Lea, S. E. (1994a). Visual perception of intentional motion. *Perception*, *23*(3), 253-268.

Dittrich, W. H., & Lea, S. E. (1994b). Visual perception of intentional motion. *Perception*, *23*(3), 253–268.

Farnioli, E., Gabiccini, M., & Bicchi, A. (2015). Optimal contact force distribution for compliant humanoid robots in whole-body loco-manipulation tasks. In *Ieee international conference on robotics and automation (icra).*

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, *21*, 1845-1853.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, *59*(2), 154-179.

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

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gordon, A. S., & Roemmele, M. (2014). An authoring tool for movies in the style of heider and simmel. In *International conference on interactive digital storytelling* (pp. 49–60).

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*(2), 243-259.

Helbing, D., & Molnar, P. (1995). Social force model for pedestrian dynamics. *Physical review E*, *51*(5), 4282.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hovaidi-Ardestani, M., Saini, N., Martinez, A. M., & Giese, M. A. (2018). Neural model for the visual recognition of animacy and social interaction. In *International conference on artificial neural networks* (pp. 168–177).

Hubbard, T. L. (2013). Launching, entraining, and representational momentum: Evidence consistent with an impetus heuristic in perception of causality. *Axiomathes*, *23*(4), 633–643.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43).

Kassin, S. (1981). Heider and simmel revisited: Causal attribution and the animated film technique. *Review of Personality and Social Psychology*, *3*, 145-169.

Kerr, W., & Cohen, P. (2010). Recognizing behaviors and the internal state of the participants. In *Proceedings of ieee 9th international conference on development and learning*.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, *21*(10), 749–759.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Michotte, A. E. (1963). *The perception of causality (t. r. miles, trans.)*. London, England: Methuen & Co. (Original work published 1946).

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning (icml)*.

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C.-C., ... Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition*, *130*, 360-379.

Peng, Y., Thurman, S., & Lu, H. (2017). Causal action: A fundamental constraint on perception and inference about body movements. *Psychological Science*, 798-807.

Sano, M., De Freitas, J., Haber, N., & Yamins, D. (2020). Learning in social environments with curious neural agents. In *42nd annual meeting of the cognitive science society (cogsci)*.

Scholl, B. J., & Tremoulet, R. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299-309.

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.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354-359.

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).

Xie, D., Shu, T., Todorovic, S., & 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.

Xie, J., Hu, W., Zhu, S.-C., & Wu, Y. N. (2015). Learning sparse frame models for natural image patterns. *International Journal of Computer Vision*, *114*(2-3), 91-112.

Zhu, S.-C., Wu, Y., & Mumford, D. (1998). Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, *27*(2), 107–126.