Partitioning the Perception of Physical and Social Events Within a Unified Psychological Space

Tianmin Shu Yujia Peng Hongjing Lu Song-Chun Zhu
{tianmin.shu, yjpeng, hongjing}@ucla.edu sczhu@stat.ucla.edu
Department of Psychology and Statistics, University of California, Los Angeles, USA

Abstract

Humans demonstrate remarkable abilities to perceive physical and social events based on very limited information (e.g., movements of a few simple geometric shapes). However, the computational mechanisms underlying intuitive physics and social perception remain unclear. In an effort to identify the key computational components, we propose a unified psychological space that reveals the partition between the perception of physical events involving inanimate objects and the perception of social events involving human interactions with other agents. This unified space consists of two prominent dimensions: an intuitive sense of whether physical laws are obeyed or violated; and an impression of whether an agent possesses intentions, as inferred from movements. We adopt a physics engine and a deep reinforcement learning model to synthesize a rich set of motion patterns. In two experiments, human judgments were used to demonstrate that the constructed psychological space successfully partitions human perception of physical versus social events.

Keywords: social perception; intuitive physics; intention; deep reinforcement learning, Heider-Simmel animations

Introduction

Imagine you are playing a multi-player video game with open or free-roaming worlds. You will encounter many physical events, such as blocks collapsing onto the ground, as well as social events, such as avatars constructing buildings or fighting each other. All these physical and social events are depicted by movements of simple geometric shapes, which suffice to generate a vivid perception of rich behavioral, including interactions between physical entities, interpersonal activities between avatars engaged in social interactions, or actions involving both humans and objects.

This type of rich perception elicited by movements within simple visual displays has been extensively studied in psychology. Prior work showed that humans possess a remarkable ability to perceive physical events and to infer physical properties (e.g., masses of objects) (Proffitt & Gilden, 1989), as well as to make causal judgment (Michotte, 1963), based on observations of the movements of two objects. Furthermore, Heider & Simmel (1944) demonstrated that humans also excel in spontaneously reconstructing social events from movements of simple geometric shapes, and describe their observations in terms of agency, goals, and social relations. These classic studies, along with a great deal of subsequent psychological research (e.g., Kassin 1981; Scholl & Tremoulet 2000; Gao et al. 2009, 2010), provide convincing evidence that human inferences about physical and social events are efficient and robust, even given very limited visual inputs.

Although many studies of both intuitive physics and social perception examined dynamic stimuli consisting of moving shapes, these research areas have largely been isolated from one another, with different theoretical approaches and experimental paradigms. In the case of physical events, research has been focused on the perception and interpretation of physical objects and their dynamics, aiming to determine whether humans use heuristics or mental simulation to reason about intuitive physics (see a recent review by Kubricht et al. (2017)). For social perception, some research has aimed to identify critical cues based on motion trajectories that determine the perception of animacy and social interactions (Dittrich & Lea, 1994; Scholl & Tremoulet, 2000; Gao et al., 2009; Shu et al., 2018). Other work focused on inferences about agents' intentions (Baker et al., 2009; Ullman et al., 2010; Pantelis et al., 2014). In contrast to the clear separation between the two research topics, human perception integrates the perception of physical and social events. Hence, it is important to develop a common computational framework applicable to both intuitive physics and social perception to advance our understandings on how humans perceive and reason about physical and social events.

In the present paper, we propose a unified framework to account for the perception of both physical events and of social events based on movements of simple shapes. We aim to construct a unified psychological space that may reveal the partition between the perception of physical events involving inanimate objects and the perception of social events involving human interactions with other agents. Specifically, we hypothesize that this unified space includes two prominent dimensions: an intuitive sense regarding whether physical laws are obeyed or violated; and an impression of whether an agent possesses intentions in the display. Note that the intuitive sense of physical violation may result from observable physical forces that can not be explained by perceived entity properties (such as motion, size, etc.) in a scene. The development of this unified space may shed light on many fundamental problems in both intuitive physics and social perception.

To construct such space, we project a video as a whole onto the space. Hence, a large range of videos can provide a distribution of observed events. We can also project individual entities in one physical or social event onto the same space, and then examine pairwise relations between the projected locations of entities in the space, which could serve as an informative cue for judging social/physical roles of entities (e.g, as an human agent or an inanimate object).

To test the hypothesized psychological space, we report experiments involving many Heider-Simmel animations in which simple moving shapes vary in degrees of physical vi-

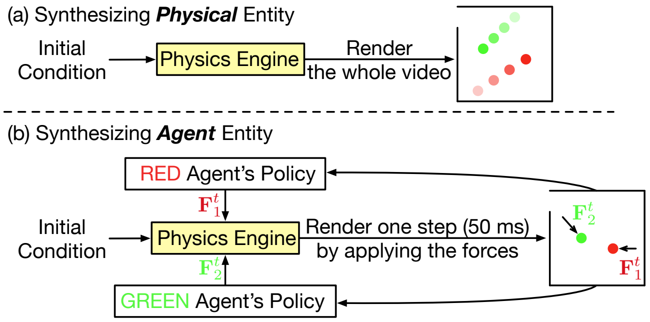


Figure 1: Overview of our joint physical-social simulation engine. For a dot instantiating a physical object, we randomly assign its initial position and velocity and then use physics engine to simulate its movements. For a dot instantiating a human agent, we use policies learned by deep reinforcement learning to guide the forces provided to the physics engine.

olation and the involvement of intention. Prior work usually created Heider-Simmel-type stimuli using manually designed interactions (Gao et al., 2009, 2010; Isik et al., 2017), rule-based behavior simulation (Kerr & Cohen, 2010; Pantelis et al., 2014), and trajectories extracted from human activities in aerial videos (Shu et al., 2018). It is challenging to manually create many motion trajectories, and to generate situations that violate physical constraints. Accordingly, we develop a joint physical-social simulation-based approach built upon a 2D physics engine (Figure 1). A similar idea has been previously instantiated in a 1D environment, Lineland (Ullman, 2015). By generating Heider-Simmel-type animations in a 2D environment with the help of deep reinforcement learning, our simulation approach is able to depict a richer set of motion patterns in animations.

This advanced simulation provides well-controlled Heider-Simmel stimuli enabling the measurement of human perception of physical and social events for hundreds of different motion patterns. We also develop general metrics to measure how well the motion patterns in an animation satisfy physics, and the likelihood that dots are agents showing intentions. These two indices were computed for each stimulus shown to human observers, allowing us to map all videos into a unified space as the two measures providing primary coordinates. In two experiments, we combined model simulations with human responses to validate the proposed psychological space.

Stimulus Synthesis

Overview

Figure 1 gives an overview of our 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[1]). If a dot represents an object, we randomly assigned the 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), but did not include friction. If a dot represents an agent, it was assigned with a clearly-defined goal (e.g., leaving 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 learned by deep reinforcement learning (see more details in a later subsection). Specifically, at each step (every 50 ms), the agent observed the current state rendered by the physics engine, and its policy determined the best force 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.

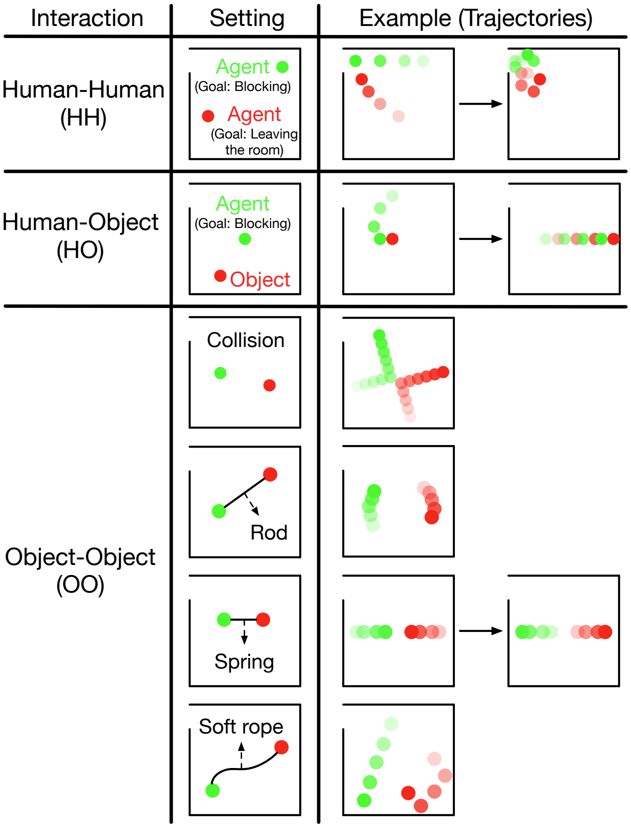


Figure 2: An illustration of three types of synthesized interactions for physical and social events. A few examples are included by showing trajectories of the two entities. The dot intensities change from low to high to denote elapsed time. Note that the connections in OO stimuli (i.e., rod, spring, and soft rope) are drawn only for illustration purpose. Such connections were invisible in the stimuli. Examples of stimuli are available at: `https://tshu.io/HeiderSimmel/CogSci19`.

Interaction Types

As summarized in Figure 2, we consider three types of interactions, including human-human (HH), human-object (HO)

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

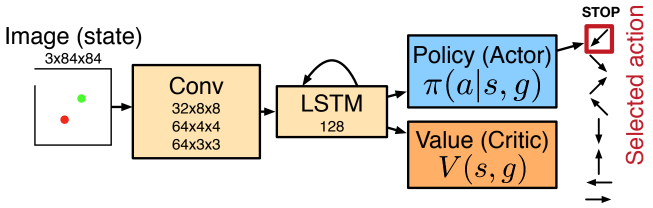


Figure 3: The deep RL network architecture for learning policy for goal-directed movements of an agent. For each goal, we train a separate network with the same architecture.

and object-object (OO) interactions, all of which are generated by the approach depicted in Figure 1. Note that in this paper we treat the terms “human” and “agent” interchangeably. When synthesizing the agents’ motion, we set two types of goals for the agents, i.e., ”leave the room” (g_1) and ”block the other entity” (g_2). Specially, in HH stimuli, one agent has a goal of leaving the room (g_1), and the other agent aims to block it (g_2); in HO stimuli, an agent always attempts to keep a moving object within the room (g_2) and the object has an initial velocity towards the door. By randomly assigning initial position and velocity to an agent, we can simulate rich behaviors that can give the impression such as blocking, chasing, attacking, pushing, etc.

In addition to the three general types of interactions, we have also created sub-categories of interactions to capture a variety of physical and social events. For OO animations, we included four events, as collision, connections with rod, spring and soft rope. Since these connections were invisible in the displays, the hidden physical relations may result in a subjective impression of animacy or social interactions between the entities. In addition, the invisible connections between objects (rod, spring, and soft rope) introduce different degrees of violation of physics in the motion of the corresponding entities if assuming the two entities are independent. For HH animations, we varied the “animacy degree” (AD) of the agents by controlling how often they exerted self-propelled forces in the animation. In general, a higher degree of animacy associates with more frequent observations about violation of physics, thus revealing self-controlled behaviors guided by the intention of an agent. The animacy manipulation introduced five sub-categories of HH stimuli with five degrees of animacy – 7%, 10%, 20%, 50%, and 100%, respectively corresponding to applying force once for every 750, 500, 250, 100, and 50 ms. In an HH animation, we assigned the same level of animacy degree to both dots.

Training Policies

As shown in Figure 1, in order to generate social events, we need sensible policies to infer the self-propelled forces for pursuing goals. However, searching for such policies in a physics engine is extremely difficult. In this study, we use deep reinforcement learning (RL) to acquire such policies, which has been shown to be a powerful tool for learning complex policies in recent studies (Silver et al., 2017).

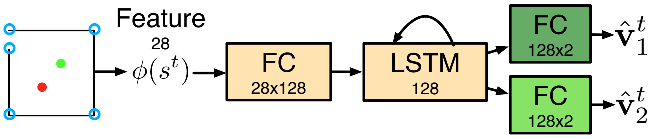


Figure 4: Network for the physical motion prediction model to emulate intuitive physics. Blue circles indicate the corners of the room used for deriving the input features.

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 3) 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 of 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 model is to train the policy network shown in Figure3 to maximize the expected return $E[\sum_{t=0}^{\infty} \gamma^t r^t]$ for each agent. The optimization was implemented using advantage actor critic (A2C) (Mnih et al., 2016) to jointly learn a policy (actor) $\pi : \mathcal{S} \times \mathcal{G} \mapsto \mathcal{A}$ which maps an agent’s state and goal to its action, and a value function (critic) $V : \mathcal{S} \mapsto \mathbb{R}$. The two functions were trained as follows (assuming that entity i is an agent):

$$\nabla_{\theta_\pi} J(\theta_\pi) = \nabla_{\theta_\pi} \log \pi(a_i^t | s_i^t, g_i; \theta_\pi) A(s_i^t, g_i), \tag{1}$$

$$\nabla_{\theta_V} J(\theta_V) = \nabla_{\theta_V} \frac{1}{2} \left(\sum_{\tau=0}^{\infty} \gamma^\tau r_i^{t+\tau} - V(s_i^t, g_i; \theta_V) \right)^2, \tag{2}$$

where $A(s_i^t, g_i) = \sum_{\tau=0}^{\infty} \gamma^\tau r_i^{t+\tau} - V(s_i^t, g_i)$ is an estimate of the advantage of current policy over the baseline $V(s_i^t, g_i)$. We set $\gamma = 0.95$ and limit the maximum number of steps in an episode to be 30 (i.e., 1.5 s). Note that we train a network for each goal with the same architecture. In HH animations, an agent’s policy depends on its opponent’s policy. To achieve a joint policy optimization for both agents, we adopt an alternating training procedure: at each iteration, we train the policy of one of the agents by fixing its opponent’s policy. In practice, we trained the polices by 3 iterations.

Inference of Physical and Social Events

Physics Inference

The first type of inference assesses the degree of violation of physics for each entity. To capture this measure, we used physical events to train a deep recurrent neural network (see

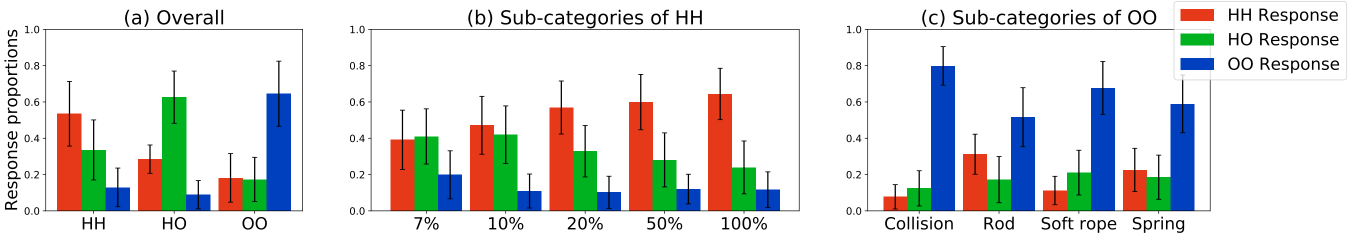


Figure 5: Human response proportions of interaction categories (a) and of the sub-categories (b,c) in Experiment 1. Error bars indicate the standard deviations across stimuli.

Figure 4) as an approximation to emulate intuitive physics. The network can predict the velocities of the two objects $\hat{\mathbf{v}}_i^t$, $i = 1,2$, given their past trajectories $\Gamma_i^t = \{s_i^\tau\}_{\tau=1}^t$. At each step, we feed a 28-dim feature vector to the network by concatenating the two dots' positions in the room, their relative positions to each other and to the five corners highlighted by the blue circles in Figure 4. We generated 2000 collision OO videos and trained the network on these videos with a 4-fold cross-validation. Using the trained network, we then conducted a step-by-step prediction of an entity's movements assuming it is an object. By comparing with the ground truth $\mathbf{v}_i^t$, we can evaluate to what degree an entity's motion is inconsistent with physics predictions:

$$D_i = \frac{1}{T}\sum_{t=1}^{T} ||\mathbf{v}_i^t - \hat{\mathbf{v}}_i^t||_2^2, \quad \forall i = 1,2. \tag{3}$$

Intention Inference

To evaluate the impression of whether a dot possesses intentions in the Heider-Simmel display, we estimate a value index (i.e., accumulated reward) from an entity's trajectory w.r.t. each possible goal. We first define a reward function:

$$R(s^t, g) = \frac{(\mathbf{x}_g^t - \mathbf{x}^t)^\top \mathbf{v}^t}{||\mathbf{x}_g^t - \mathbf{x}^t||_2 || \cdot \mathbf{v}^t||_2}, \tag{4}$$

where $\mathbf{x}^t$ and $\mathbf{v}^t$ are the position and velocity of an entity extracted from its state s^t, and $\mathbf{x}_g^t$ is the position of the goal. For "leaving the room", $\mathbf{x}_g^t$ is the door's position, whereas $\mathbf{x}_g^t$ denotes the position of the other entity for "blocking". Intuitively, this reward function evaluates whether the entity is moving towards certain goal locations. Consequently, we can compute the overall value by selecting the most likely goal:

$$V_i = \left[\max_{g \in \mathcal{G}} \frac{1}{T}\sum_{t=1}^{T} R(s_i^t, g)\right]_+, \quad \forall i = 1,2, \tag{5}$$

where $[x]_+ = \max(x, 0)$. Note that V_i defined here is different from the one in Eq. 2. Ranging from 0 to 1, a higher value of V_i indicates that the entity i shows a clearer intention and is more likely to be an agent. We remove the moments when the denominator in Eq. (4) is too small for the robustness of the value estimate. Considering the complexity of optimal planning in the continuous physical environment, the proposed value index offers a simplified measure of goal inference by inverse planning (Baker et al., 2009; Ullman et al., 2010).

Experiment 1

Participants

30 participants (mean age = 20.9; 19 female) were recruited from UCLA Psychology Department Subject Pool. All participants had normal or corrected-to-normal vision. Participants provided written consent via a preliminary online survey in accordance with the UCLA Institutional Review Board and were compensated with course credit.

Stimuli and Procedure

850 videos of Heider-Simmel animations were generated from our synthesis algorithm described above, with 500 HH videos (100 videos for each AD level), 150 HO videos, and 200 OO videos (50 videos for each sub-category). Videos lasted from 1 s to 1.5 s with a frame rate of 20 fps. By setting appropriate initial velocities, the average speeds of dots in OO videos were controlled to be the same as the average speeds of dots in HH with 100% ADs (44 pixel/s). 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 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. Your task is to 'identify' these two dots based on their movement. There are three possible scenarios: human-human, human-object, or 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.

Results

Human response proportions are summarized in Figure 5. Response proportion of human-human interaction swas significantly greater than the chance level 0.33 ($t(499) = 25.713$, $p < .001$). For HO animations, response proportion of human-object interaction was significantly greater than the other two responses ($p < .001$). Similarly, response proportion of object-object was greater than the other two responses ($p < .001$) for OO animations. These results reveal that human participants identified the main characteristics of different interaction types based on dot movements.

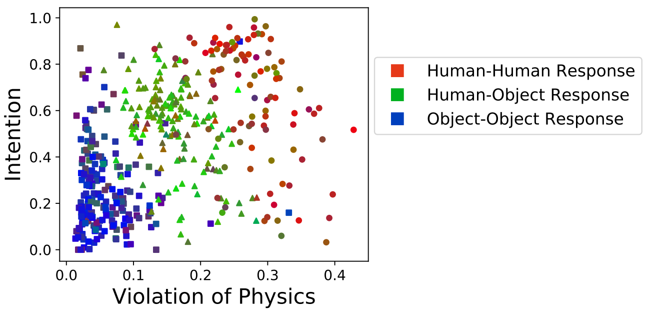


Figure 6: Constructed psychological space including HH animations with 100% animacy degree, HO animations, and OO animations. In this figure, a stimulus is depicted by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. The two coordinates of the space are the averaged measures between the two entities, as the measure of the degree of violation of physical laws (horizontal) and the measure of values indicating the presence of intention. 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).

Next, we examined human responses to the sub-categories within the HH and OO animations. We first used the animacy degree as a continuous variable and tested its effect on human responses in the HH animations. With increases in degree of animacy in HH, the response proportion of human-human interaction increased significantly as revealed by a positive correlation ($r = .42$, $p < .001$). This finding suggests that humans are sensitive to the animacy manipulation in terms of the frequency with which self-propelled forces occurred in the stimuli. For the OO animations, the response proportion for object-object interaction among the four sub-categories yielded significant differences ($F(3,196) = 34.42$, $p < .001$ by an ANOVA), with the most object-object responses in the collision condition, and the least in the rod condition. Pairwise comparisons among the four-categories show significant difference between collision and everything else ($p < .001$), between soft rope and rope ($p < .001$), and also between soft rope and string ($p = .018$); there is a marginally significant difference between rod and string ($p = .079$).

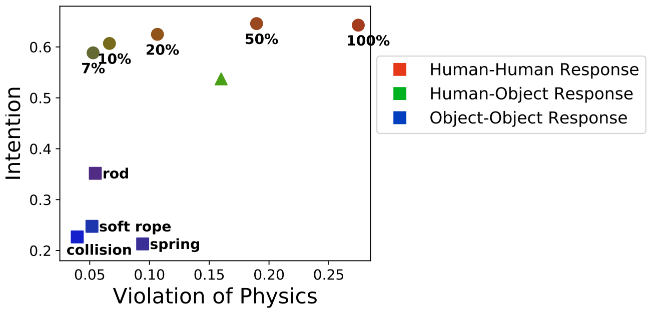


Figure 7: Centers of all types of stimuli.

We then combined human responses and the model-derived measures for each animation stimulus to depict the unified psychology space for the perception of physical and social events. Figure 6 presents the distributions of 100 HH videos with 100% animacy degree, 150 HO videos, and 200 OO videos, all in this unified space. In this figure, an animation video is indicated by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. Specifically, the values of its RGB channels are determined by the average human-human responses in red, human-object responses in green, and object-object responses in blue. The mark shapes of data points correspond to the interaction type used in the simulation for generating the synthesized animations. The coordinates of each data point were calculated as the model-derived measures averaged across the two entities in an animation. The resulting space showed clear separations between the animations that were judged as three different types of interactions. Animations with more human-human interaction responses (red marks) clustered at the top-right corner, corresponding to great values of intention and strong evidence signaling the violation of physics. Animations with high responses for object-object interactions (blue marks), located at the bottom left of the space, show low values of intention index and little evidence of violation of physics. Animations with high responses for human-object interactions (green marks) fell in the middle of the space.

To quantitatively evaluate how well the model-derived space accounts for human judgments, we trained a classifier using the coordinates derived in the space shown in Figure 6 as input features (D and V for the indices of physical violation and intention respectively). For each ground-truth type of interactions $k \in \{\mathrm{HH}, \mathrm{HO}, \mathrm{OO}\}$, we fit a 2D Gaussian distribution $p_k(D,V)$, using half of the stimuli as training data. Then for a given animation with the coordinates of (D,V), the classifier predicts $p(k|D,V) = \frac{p_k(D,V)}{\sum_k p_k(D,V)}$ for animations in the remaining half of the stimuli. The correlation between the model predictions and average human responses was 0.748 ($p < .001$) based on 2-fold cross-validation. Using a split-half reliability method, human participants showed an inter-subject correlation of 0.728 ($p < .001$). Hence, the response correlation between model and humans closely matched inter-subject correlations, suggesting a good fit of the unified space as a generic account of human perception of physical and social events based on movements of simple shapes.

We examined the impact of different degrees of animacy on the perception of social events, and how different subcategories of physical events affect human judgments on interaction types. The unified space provides a platform to compare these fine-grained judgments. Figure 7 shows the centers of the coordinates and the average responses for each of the sub-categories. We first found that, with a decreased degree of animacy, the intention index in HH animations was gradually reduced towards the level of HO animations. Meanwhile, human judgments of these stimuli varying from low to high degree of animacy transited gradually from human-object responses to human-human responses, consistent with the trend that the data points moved along the physics axis. Among all physical events, the rod and spring conditions showed the highest intention index and the strongest physical violation, respectively, resulting in a greater portion of human-human interaction responses than the other categories.

Experiment 2

In Experiment 1, human participants were asked to classify the three interaction types. But for human-object responses, the assignment of the roles to individual entities was not measured. In Experiment 2, we focused on stimuli that elicited the classification of human-object responses, and asked participants to report which dot was a human agent, and which dot was an inanimate object. Specifically, the role assignment in the human-object responses helps us identify some key characteristics in the psychological space that signal a human-object interaction.

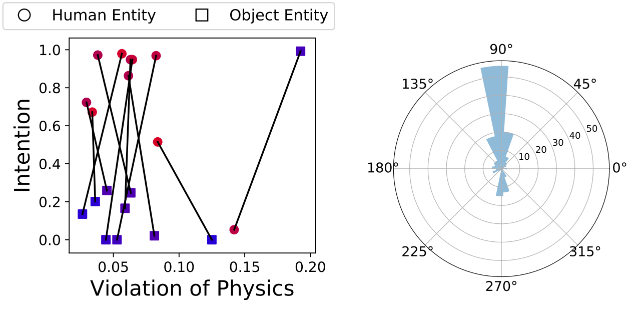
Methods

25 participants (mean age = 21.3; 19 female) were recruited from the UCLA Psychology Department Subject Pool. 216 videos were selected from Experiment 1 based on the criterion that more than 40% of subjects judged the HH videos or OO videos as human-object interaction. 201 videos were HH videos and the other 15 were OO videos.

The procedure was the same as Experiment 1 except that on each trial, subjects were asked to complete two tasks: first to judge the interaction type; then if the judgment was human-object, they were further asked to report which dot represented a human agent and which dot represented an object.

Results

We projected all entities onto the psychological space based on the model-derived measures for each individual entity, and connected a pair of the two entities that appeared in the same video. We visualized 10 animations that yielded high human-object response proportions and the most consistent role judgment among participants as shown in Figure 8a, where circles represent the dots that were frequently identified as humans, and squares represent the dots identified as objects. The resulting segments showed a common feature in that the connection of the two entities in the space depicted a near-vertical orientation, primarily due to high intention value for the human dot, and low intention value for the object dot. To further examine the orientations in the space for the human-object responses, we calculated the histogram of the orientations for animations judged as human-object interactions, which shows a high concentration around 90 degrees (see Figure 8b). This finding suggests that the two dots in the Heider-Simmel animations elicited similar degrees of physical violation, but one of them showed a much clearer intention. Note that this analysis excluded 38 stimuli in which participants did not show consistency in the role judgment (each entity was judged as a human or an object by exactly half of the participants).



(a)

(b)

Figure 8: Human and model-simulation results in Experiment 2. (a) Representative cases of animations that elicited the human-object responses, located in the space with model-derived coordinates. The colors reflects average human responses of assigning a dot to the human role (red) and to the object role (blue). (b) Orientation histogram of the segments connected by the concurrent pairs of entities in an animation.

Conclusion

In this study, we propose a unified psychological space to account for human perception of physical and social events from movements of simple shapes in Heider-Simmel animations. The space consists of two primary dimensions: the intuitive sense of violation of physics, and the impression of intentions. We tested the space by measuring human responses when viewing a range of synthesized stimuli depicting human-human, human-object, and object-object interactions in the style of Heider-Simmel animations. We found that the constructed physics-intention space revealed clear separations between social and physical events as judged by humans. Furthermore, we trained a classification model based on the coordinates of each stimulus in this space. The resulting model was able to predict human classification responses at the same level as human inter-subject reliability.

The present paper provides a proof of concept that the perception of physical events and social events can be integrated within a unified space. Such common representation enables the development of a comprehensive computational model of how humans perceive and reason about physical and social scenes. Perhaps the most surprising finding in our work is that the classification result based on just the two measures reflecting the violation of physical laws and the estimate of intention can predict human judgment very well, reaching the same level as inter-subject correlation. The good fit to human responses across a range of Heider-Simmel stimuli demonstrates the great potential of using a unified space to study the transition from intuitive physics to social perception.

The main benefit of constructing this psychological space is to provide an intuitive assessment for general impressions of physical and social events. To build up such representation, humans or a computation model may use various cues to detect intentions and/or physical violations; such cue-based detection is usually subjected to personal preferences. Instead of discovering a list of cues for distinguishing between physical events and social events, the proposed space offers an abstract framework for gauging how humans' intuitive senses of physics and intentions interplay in their perception of physical and social events.

This work provides a first step toward developing a unified computational theory to connect human perception and reasoning for both physical and social environments. However, the model has limitations. For example, the simulations are limited by a small set of goals, and the model requires predefined goals and good knowledge about the constrained physical environment. Future work should aim to extend the analysis to a variety of goals in social events (Thurman & Lu, 2014), to develop better goal inference, and to support causal perception in human actions (Peng et al., 2017). 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, we have only examined human perception of physical and social events on short stimuli with only two entities. Generating longer stimuli with more entities and analyzing human perception on them will further help reveal the mechanisms underlying humans' physical and social perception.

Acknowledgement

We thank Ciaran Zhou, Claire Locke, Huiwen Duan, Suhwan Choi, and Zhibo Zhang for assistance in data collection. We also thank Dr. Tao Gao at UCLA 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.

References

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

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

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.

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

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.

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*, 360379.

Peng, Y., Thurman, S., & Lu, H. (2017). Causal action: A fundamental constraint on perception and inference about body movements. *Psychological Science*, 0956797617697739.

Proffitt, D. R., & Gilden, D. L. (1989). Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 384-393.

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.

Thurman, S. M., & Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PloS one*, *9*(11), e112539.

Ullman, T. D. (2015). *On the nature and origin of intuitive theories: Learning, physics and psychology*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

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).