Generalizing physical prediction by composing forces and objects

Haoliang Wang
Department of Psychology
UC San Diego
haw027@ucsd.edu

Kelsey Allen
DeepMind
krallen@google.com

Edward Vul
Department of Psychology
UC San Diego
evul@ucsd.edu

Judith E. Fan
Department of Psychology
UC San Diego
jefan@ucsd.edu

Abstract

Our ability to make reliable physical predictions even in novel settings is a hallmark of human intelligence. Here we investigate how people infer multiple physical variables simultaneously and compose them to generalize to a novel scenario. Participants (N=203) observed a series of balls launched at different angles in a 2D virtual environment and generated predictions about their trajectories. We found that people could infer the masses of different balls based on these observations, as well as the existence of a latent "wind" force, and compose knowledge of these two variables to generalize to novel situations in a subsequent test phase. We modeled this generalization as the consequence of being able to simulate trajectories by independently combining force and mass information in accordance with Newtonian mechanics. To validate this approach, we also tested several alternative models and compared their generalization behavior to one another and to that of people. Together, our study points to the value of using generalization to probe the underlying representations supporting physical prediction.

Keywords: intuitive physics; world model; compositional generalization; computational models

Introduction

People readily make physical predictions about how objects will behave even in novel situations. For example, golfers can use their prior knowledge about golf balls and wind forces to play on a windy day; gamers playing *Super Mario Bros.* quickly learn how different characters react differently to the unnatural gravity in the game, and later readily control these avatars to accomplish tasks in underwater scenarios by factoring in water resistance. Indeed, living in an uncertain and open-ended physical world, a fundamental goal of our cognition is to generalize from limited experience so as to behave appropriately in unpredictable future tasks and situations. How do people learn to "carve physics at its joints" – that is, to uncover hidden variables and rules that can be flexibly used to generalize to new scenarios?

One possibility is that people are not just learning to map input sensory information to output predictions, but are rather inferring the latent properties in a structured generative world model – an internal model encoding the physical dynamics of how the world works (Battaglia, Hamrick, & Tenenbaum, 2013). Prior work in intuitive physics has established that people can infer latent physical parameters like mass (Sanborn, Mansinghka, & Griffiths, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016) and friction (Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018) by observing object's motion. In particular, it has been argued that people's inference and judgements about physical properties can be explained by having a noisy Newtonian internal physics model (Ullman, Spelke, Battaglia, & Tenenbaum, 2017). Going beyond just inferring a single *parameter* in the physics model, it has also been found that people are able to simultaneously induce the conceptualization of objects as well as the *causal relationships* between them by watching objects interact with each other in the domain of magnetism (Bonawitz, Ullman, Gopnik, & Tenenbaum, 2012). Together, these findings suggest that people can learn an internal world model that encodes the underlying dynamics of the physical world at multiple levels: from underlying causal structure to specific parameters.

However, if world models could not extend to novel objects and situations, they would be of limited use to us. Therefore, a crucial aspect of learning a structured world model is that people should be able to flexibly compose the variables in the model to make reliable physical predictions when faced with *novel* scenarios that are related but nonidentical to past experiences. To date, however, few studies have investigated how or whether people are able to accomplish this.

In this paper, we sought to explore how people learn physical world models such that they can compositionally generalize to novel scenarios and make reliable predictions. We focus on a *specific* kind of physical world model, namely models that encode the latent forces and masses of objects in an environment. This kind of world model, although simple, can have a wide range of variations (e.g. types of forces, different mass values) and is a prerequisite for learning more complex world models. To this end, we developed a novel paradigm where participants must infer multiple latent variables of the physical dynamics during training and compose them to generalize in the test phase. Specifically, we ask participants to play a physics-based video game. In this game, participants use a paddle to catch three balls of different masses in two environments where different latent forces (downward gravity and a wind force blowing to the right) are at play. People were trained on 5 out of these 6 ball-environment combinations and then asked to generalize to the held-out combination. In order to succeed at this task, participants must infer the latent structure (e.g. the existence of different latent forces in different environments) as well as physical parameters (e.g. mass of different balls) of the underlying dynamics and compose

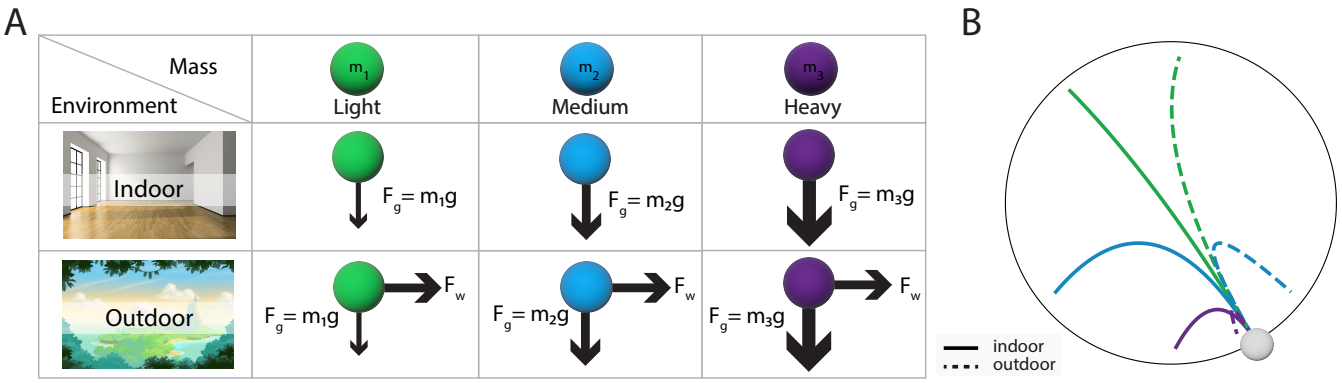

Figure 1: (A) The 2×3 design matrix of our experiment, where participants were trained on 5 out of these 6 cells, and asked to generalize to the held-out cell. The choice of held-out cell was counterbalanced across participants. (B) Different trajectories of a ball when its mass and the environment varies.

these to generalize in the novel ball-environment combination. We find that people can learn both latent variables, and critically compose this knowledge to generalize to the novel combinations in the test phase. Our modeling results suggest that people achieve such generalization, in part, by constructing composable internal models of the physical scene and performing model-based compositional generalization.

Experiment

Participants

203 participants (100 female; mean age = 25.9 years) recruited from Prolific completed the experiment. Data from all participants were included as all met our preregistered inclusion criteria. Participants provided informed consent in accordance with the UC San Diego IRB. The experiment lasted approximately 35 minutes and participants were paid $14/hr based on this expected completion time.

Task environment & procedure

To probe physical prediction, in this experiment, we ask participants to play a virtual game of catch. A ball is launched from a point on a large circle, and the participants' task is to move a rectangular paddle along the outside of the circle to catch the ball (Figure 1B). Each trial began with the paddle placed at 3 o'clock, participants then adjusted the paddle's location with the arrow keys. When participants were satisfied with the paddle's location, they launched the ball using the spacebar (as soon as the ball was launched, they could no longer adjust the paddle location). The ball's launch trajectory was animated. If the ball made contact with any part of the paddle, this was considered a success. Participants then pressed the spacebar to proceed to the next trial. We manipulate the following variables in each trial: the environment where the participants perform this task, the mass of the ball, the location where the ball was launched, and the force with which the ball was launched.

In order to introduce different latent forces that require different predictions to maintain high accuracy, we use two environments cued by different background images. In one environment, there is only gravity (F_g) pulling downward; and in the other environment, there is both a downward gravity force and a rightward wind force (F_w). As these forces are evocative of indoor/outdoor environments, we use the indoor/outdoor nomenclature for simplicity throughout the paper. To elicit participants' inferences about physical parameters, we use three types of balls: light, medium and heavy. All balls are the same size, but have different colors and textures, allowing participants to learn a color/texture $\rightarrow$ mass mapping throughout the experiment. The correspondence between the color/texture of the ball and its mass is shuffled across participants. As a way of measuring how well people could make predictions under different physical conditions, the ball appears at a location sampled from each of the 12 hours on a clock face, and is launched towards the center of the big circle with an initial force whose direction and magnitude were indicated by an arrow, either strong (red) or soft (orange). We manipulate mass (light, medium, heavy) and environment (indoor, outdoor) using a "2×3 factorial design" such that succeeding on any given trial required combining these two latent variables (see Figure 1A). Each ball-environment combination consists of 24 trials (12 launching locations $\times$ 2 launching forces).

The game consists of a training phase and test phase. In the training phase, participants are exposed to five of the six ball-environment combinations. The subsequent test phase only includes trials with the remaining ball-environment combination. To generalize to the test phase, the participants need to successfully infer the underlying structure (the existence of gravity/wind) as well as the specific parameters (how strong the gravity/wind is, and how heavy the balls are) of the physical environment. We randomly assign participants to each of six groups defined by which ball-environment combination was used at test. To give participants an opportunity to observe how each ball behaved under different launch conditions (launching location, launching forces) in the same environment, we divided the 120 (24×5) training trials into

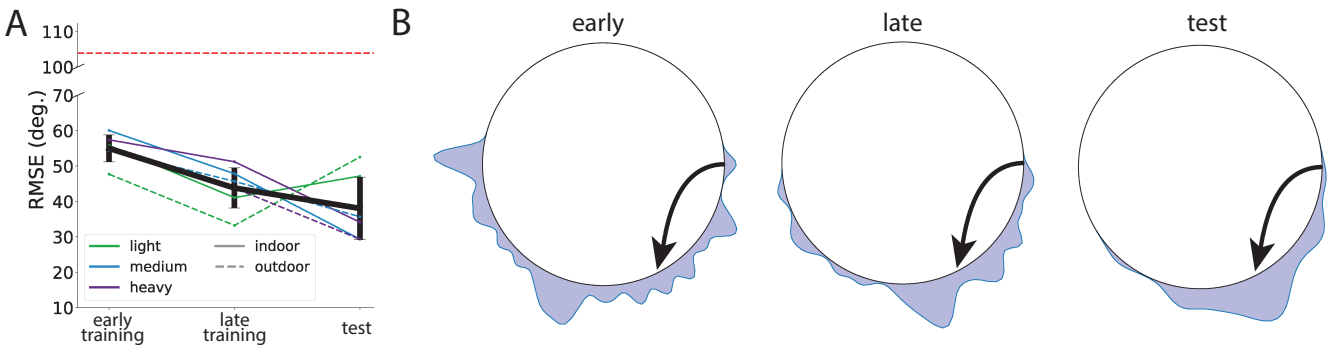

Figure 2: (A) RMSE for all 6 conditions, black thick line shows the mean and standard deviation. Dashed red line represents expected performance under random guessing (Rinaman et al., 1996). (B) Three trials with the same launching condition were selected from the three timepoints. Black-dotted trajectory demonstrates the movement of the ball in each trial. When first playing the game, participants displayed high bias and variance placing the paddle (left most trial), resulting in high RMSE; in the test phase, both bias and variance have shrinked dramatically (right most trial).

4 blocks by environment (e.g. indoor first, then two outdoor blocks, then indoor). Transitions between blocks were not marked, and the order in which participants encountered the indoor/outdoor environments was counterbalanced across participants. Within each block, we randomized the sequence of launching location, launching force and ball mass.

Results

People can learn the dynamics over time

Given that participants had no prior exposure to this task environment, we first sought to evaluate how accurate participants' predictions were in absolute terms. On each trial, we measured the participants' paddle location, the ball's ground truth landing location when it crossed the large circle, and the angular difference between them. To quantify accuracy of participants' behavior, the root average squared deviation from the ground truth landing location in degrees was analyzed (root mean squared error, RMSE). We calculated RMSE for the first and second half of training, and test phase, collapsing over the feature dimensions that varied (launching force, launching location, ball mass, environment) because the design was carefully counterbalanced such that each feature was equally likely to be practiced. Figure 2A shows RMSE for all 6 conditions. Participants' performance was significantly above chance at every point during this experiment ($t = -75.16$, $p < 0.001$). Initially, RMSE was high (mean=55.01°), presumably reflecting the fact that participants were uncertain about the physical dynamics when they were first introduced to this task context; participants would have faced high error when their estimates of either the structure (e.g. the existence of wind in the outdoor environment) or the parameter (e.g. the mass of the balls, the magnitude of the wind, etc.) was wrong. Figure 1B shows an example of how different estimates lead to very different predictions of the ball's landing location. By the end of the experiment, however, participants significantly improved (mean=38.05°; $b = -11.12, t = -4.76$, $p < 0.001$). Different conditions showed similarly low error rates, with the exception of the lightest ball in the outdoor environment, reflecting the fact that the lightest ball's behavior is relatively hard to predict when wind is at play because the amount it accelerates due to the wind is relatively high compared to the heavier balls. Qualitatively, 2B shows the distribution of participant paddle placements in the first half of training trials (early), second half (late) and test condition (test) as a histogram. Broadly, this suggests that while people may have struggled to learn the mechanics of the task at the beginning, they rapidly improved over time.

Model-based generalization can account for test phase behavior

In the last section, we observed that test phase performance was as good or better than in the training phase, despite test phase trials consisting of novel combinations of indoor/outdoor context and ball mass. What might account for such behavior? One possibility is that from the observations in the training phase, participants successfully learn a world model encoding the latent forces of the different environments and masses of the balls, enabling them to compose the two pieces of information during the test phase to predict the ball's trajectory and place their paddle accordingly.

If this is the case, that is, if the participants' behavior is in accordance with their world model's prediction, then from their paddle placement, we should be able to work backwards and infer the world model they have in their mind. To test this hypothesis, we adopted Bayesian inference to search for the best explaining world model given the participants' paddle placement:

$$P(M|D) = \frac{P(D|M)P(M)}{\sum_{M'} P(D|M')P(M')} \quad (1)$$

where D stands for participants' data (paddle placements), M for participants' mental world model, $P(D|M)$ for the likelihood of the participants' paddle placements given a hypothesized world model, and $P(M)$ for the prior on models.

As mentioned before, participants need to infer the existence of latent forces (F_g and F_w) and estimate the mass of different balls (m_1, m_2 and m_3) to succeed in this task. Since the gravitational force is constant in all contexts throughout the task, we only infer mass parameters and F_w. It is worth noting that although minimal, this hypothesis space encompasses a large variety of world models that participants may have. For example, if the wind magnitude were 0 in a participant's world model, they would think there is only downward gravity and wind does not exist, which is the correct model for the indoor environment. By varying the ball mass parameter in a participant's world model, they would have very different predictions as to where a given ball would land in the same environment (see Figure 1B for the trajectories and landing locations of the same ball under different world models).

Given an initial launching force and launching location on the circle, the participants' world model M can simulate the trajectory of the ball, and estimate the balls' subsequent landing location when it crosses the large circle. To capture participants' motor noise when placing the paddle, we use a wrapped normal distribution (defined over angles around a circle) for the likelihood term $P(D|M)$: $P(D|M) = \mathcal{N}(\mu, \sigma)$, where the mean μ is the estimated landing location using model M, and σ indicates how noisy the participants are when sampling a paddle placement given the estimated landing location. A uniform prior $P(M)$ was used for wind (F_w) (in ranges $[-50, 100]$), mass (m) (in ranges $[0.5, 5]$), and σ (in ranges $[0.05, \pi/2]$).

We use participants' paddle placement during the test phase as D to measure their estimation of variables in a novel scenario. We perform a grid search to obtain the posterior distribution and use the posterior mean as an estimate for each participant's (F_w, m, σ). Figure 3 shows the estimated F_w and m for each participant for all six conditions. The posterior means and standard deviations for each variable are shown as the colored crosses. The true underlying model parameters of the three masses ($m1$, $m2$ and $m3$) and wind force (F_w) are shown as red crosses. On the whole, human estimates appear to track ground truth parameters quite well, although there is some evidence of shrinkage, or regularization: the lightest object mass is overestimated by about 10%, and the heaviest object's mass is underestimated by about 10%. This pattern is consistent with shrinkage due to hierarchical inference in the face of uncertainty (Gelman & Hill, 2006).

People are consistent on compositional generalization, regardless of their performance

Our model-based analysis revealed substantial variation between individuals (Figure 3), with some participants closer to ground truth during generalization and others farther away. For these people who are farther away from ground truth during generalization, is it because they have learned the right world model but failed to appropriately combine the information, or is it because they are slightly off when inferring the latent physical properties during training but are still able to combine these properties when generalizing (even if they are not veridical)?

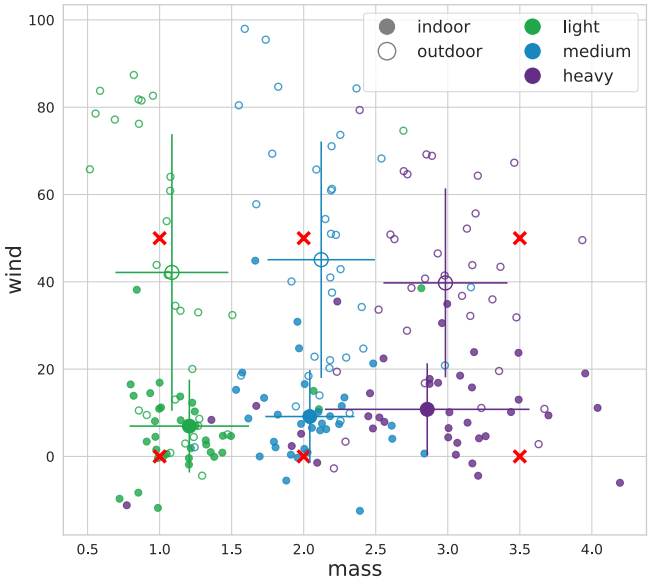

Figure 3: Fitted parameters for each individual for all conditions, wind (F_w) on the y axis and mass (m) on the x axis. Error bars represent standard deviation on each dimension. Red crosses indicates the ground truth wind and mass values for the 6 conditions.

One way to tease apart these hypotheses is to see whether the estimated wind and mass values are consistent across the training and test phase. To this end, for each condition, we compare the estimated wind and mass values for both the training and test phase. For example, if a participant was asked to generalize to the medium ball in the outdoor environment, we analyzed all the trials containing either the medium ball, or in the outdoor environment (see Figure 1, these trials correspond to the cells in the same row or column as the test phase cell in the design matrix). Adopting the same Bayesian method described in the previous section, we then estimate the wind magnitude and ball mass using the paddle placements in these trials as data D. We compute the correlation between these estimated values and those estimated using the test phase. Though the estimated F_w and m in the training phase are noisier because they include trials spanning the entire training phase (people's world model are noisy and uncertain at the beginning and gradually improve, see Figure 2), we still see reliable correlations between the values estimated during training and generalization (wind: $r = 0.65$, $p < 0.001$; mass: $r = 0.67$, $p < 0.001$). These results suggest that people are internally consistent between the training and test phases, even when their estimate of either a ball's mass or the wind was not veridical.

Comparing different computational models to human behavior

Our results so far suggest that participants are able to learn a mental world model from experience and compose their understanding such that they can generalize to unseen scenarios.

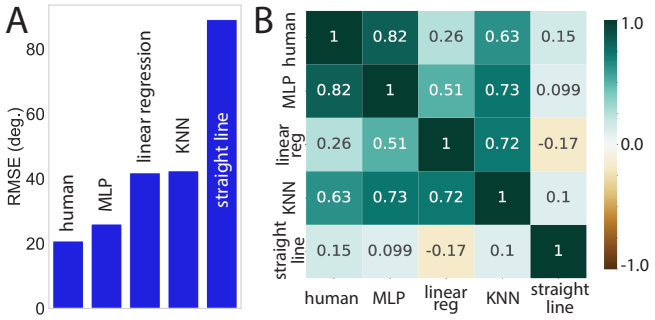

Figure 4: (A) RMSE of average response for models and human. (B) Correlation between different model and human's signed errors.

In this section, we explore several *alternative* computational models that make different assumptions about the underlying representation used to drive decisions, and compare their predictions to human behavior.

To this end, we designed and implemented several classes of computational models as possible alternative accounts for how participants might perform this task. We use the same input for every model: on every trial, launching force, ball color and environment are encoded as categorical variables (one-hot); with launching location as a numeric value.

- **Straight line heuristic**: One possible account is that people are using a simple heuristic that assumes there is no force at play and objects always travel in straight lines. If this were true, since balls are always launched towards the center of the large circle, participants would always place the paddle across the circle. This heuristic is accurate when the ball is launched from 12 o'clock in the indoor environment where there is only gravity.

- **Linear regression**: This models assumes that its input and the ball's landing location follow a linear relationship $location = \sum_i w_i \times var_i + b_i$. The free parameters are its coefficients w_i and bias b_i for each input variable var_i, which are fitted using least squares.

- **Memory retrieval**: When asked to predict on a new trial, this model searches through the trials it has already played before to find the K most similar trials in terms of input, and then averages the landing locations on these trials to make a prediction. We implemented a K-Nearest Neighbor (KNN) model for this, and used Manhattan distance for the categorical variables in the input, and angular distance ($L2$) for launching location. The free parameters are K and the relative weighting of the angular distance compared to the Manhattan distance for the similarity calculation.

- **MLP**: We implemented a 4-layer (200 and 100 units for the two hidden layers respectively) fully connected neural network, using ReLU as its activation function and stochastic gradient descent (SGD) optimizer with a learning rate of 0.2. The free parameters are the weights and biases of the connections.

We take the "best performing variant" of each class by optimizing their free parameters (except for the straight line model which has no free parameters) using the ball's ground truth landing locations in the training trials, and ask them to predict on the generalization trials.

To systematically compare the *pattern* of errors made by the models and humans, we run each model multiple times to get a distribution of predictions for each trial. The straight line heuristic model and the linear regression model are deterministic, thus for each trial we only have one prediction from each of these two models. The variation in the MLP model comes from running with different seeds; for the memory retrieval model, if two neighbors have identical distances but different predicted landing locations, the result will depend on the ordering of the training data, resulting in a distribution of paddle placements.

For each model, we calculate RMSE using the averaged predictions on each trial. Figure 4A shows RMSE of the models compared to humans. The straight line heuristic performs worst at capturing human behavioral patterns, providing strong evidence against the possibility that participants simply placed their paddle across the circle. The linear regression model and KNN performed about as well as one other, but reliably worse than humans. Consistent with our hypothesis above, this indicates that when performing the task, participants were not retrieving and averaging exemplars from memory nor fitting a straight line mapping the input to output. Of the four models, the MLP performs the best in terms of RMSE. We further calculated the correlations between each model and human's signed errors, as shown in Figure 4B, defined as the signed angular deviation between human/model's prediction and the ground truth landing location. Positive errors mean that human/model's prediction is more "counterclockwise" than the ground truth and negative for clockwise.

Quantitatively, the MLP outperforms the other alternative models in capturing human behavior. However, we also discovered systematic deviations between how the MLP and people behaved on test trials. In Figure 5, we show the same ball launched from four different locations that spanned the circle (2 o'clock, 10 o'clock, 6 o'clock and 12 o'clock) in the indoor environment. Notice that participants' responses (shown in blue) are almost always centered at the ground truth landing location, whereas different models have different patterns of biases. Perhaps most interestingly, when the ball was launched from 6 o'clock, it went straight up but fell back before reaching the top of the circle. Participants who correctly inferred the ball's mass would place their paddle at the bottom, but for those who inferred a lighter mass, they would place the paddle atop. This is indeed what happened (see the two modes of participants' response when the ball was launched from 6 o'clock in Figure 5). This, again, indicates that some participants are slightly off when inferring the parameters, but at the same time provides strong evidence that participants do have a mental model of the world which en-

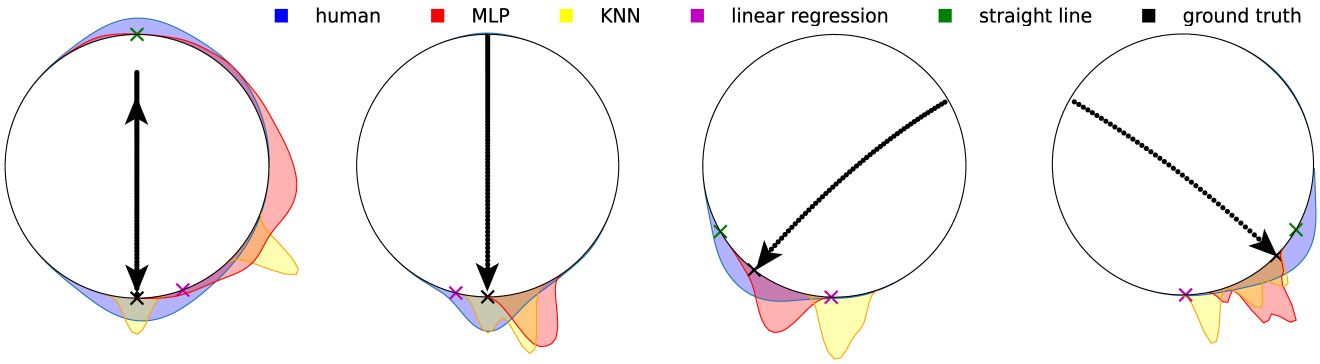

Figure 5: Four trials selected from the test phase. In these trials, the ball was launched from 6 o'clock, 12 o'clock, 2 o'clock and 10 o'clock in the indoor environment respectively. MLP, KNN and human predictions are shown in distribution. Straight line model and linear regression model's predictions are shown by the $\times$ marker because they are deterministic (see text). We also show the ground truth trajectory and landing location of the ball in each plot.

codes the notion of "mass", as opposed to using some heuristics to predict where the ball would land. The MLP in this case, however, predicts that the ball would swerve to the right. One possible explanation would be that it does not infer an explicit notion of mass but is rather "averaging" responses in the training data to generalize to this new ball and environment. When the ball was launched from 12 o'clock, it went straight down due to gravity. This is probably the easiest trial of all 144 trials, which is also verified by participants' concentrated predictions at 6 o'clock shown in the figure, indicating that they have inferred the correct latent force model for the indoor environment during the training trials and successfully applied that to this novel ball-environment combination during generalization. For this very simple trial however, the MLP systematically deviates from the ground truth and human predictions, suggesting that it has not learned a composable world model to generalize. Broadly, this suggests that none of these alternative models provides a satisfying account of how humans were able to perform this task.

Discussion

How are people able to learn the underlying dynamics of the physical world and compositionally generalize? We developed a novel paradigm where participants must infer multiple latent variables of the physical dynamics and compose them to generalize in a novel scenario. We found that people can learn these variables simultaneously over training, but also compose novel combinations of ball masses and wind conditions at test. A variety of alternative models fail to capture the same pattern of results seen in people, suggesting that people are using compositional model-based generalization to solve the task.

A key question raised by this paper concerns the representation and learning mechanism that underlie such flexible generalization. It is possible that structured representation may play a crucial role, and understanding how humans acquire such rich world models may be critical for developing AI agents that learn and generalize as flexibly as humans do. In future work, we plan to investigate how structured computational models can account for human behavior. One possible direction is using probabilistic programs as the representation of world models (Lake, Salakhutdinov, & Tenenbaum, 2015). The compositional nature of programs naturally lends itself to modeling how people compose knowledge of different variables in the physical world. Acquiring world models thus becomes program synthesis (Gulwani, Polozov, Singh, et al., 2017). Human error patterns might be reproduced by imposing uncertainty on the wind and mass variables in such programs. Future work should also further develop more substantial tests of generalization and continual learning in physics that can be used to more strongly distinguish between different models. Evaluations in these benchmarks will be critical to expose the extent to which current state-of-the-art algorithms for physical reasoning emulate human behavior in this domain, as well as potential gaps for future algorithms to fill (Bear et al., 2021).

In this paper, we found that participants could successfully compose their knowledge about mass and latent forces to generalize to a novel scenario, it is less clear, however, whether they were able to distinguish the functional form of these two forces: gravity ($F_g = mg$) is mass dependent, but wind ($F_w = F_0$) is not. In future work, we plan to manipulate the functional forms of latent forces and evaluate participants' generalization behavior by probing physical predictions. Furthermore, people's world models are likely to encode more information than just object masses and latent forces (Ullman & Tenenbaum, 2020), an important direction for future work is to investigate how people are able to acquire more complex physical world models and the role they play in generalization.

In sum, our paper reveals novel insights about how people learn about the underlying dynamics of the physical environment as well as how their world models might be structured. In the long term, such studies may shed light on the inductive biases as well as learning mechanisms that enable rapid learning and flexible generalization seen in humans.

Acknowledgments

We would like to thank Nadia Polikarpova and members of the Cognitive Tools Lab at UC San Diego for helpful discussion. This work was supported by NSF CAREER Award #2047191 and an ONR Science of Autonomy Award to J.E.F.

References

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.

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., ... others (2021). Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

Bonawitz, E., Ullman, T., Gopnik, A., & Tenenbaum, J. (2012). Sticking to the evidence? a computational and behavioral case study of micro-theory change in the domain of magnetism. In *2012 ieee international conference on development and learning and epigenetic robotics (icdl)* (pp. 1–6).

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gulwani, S., Polozov, O., Singh, R., et al. (2017). Program synthesis. *Foundations and Trends® in Programming Languages*, *4*(1-2), 1–119.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Rinaman, W., Heil, C., Strauss, M., Mascagni, M., & Sousa, M. (1996). Probability and statistics. *Standard mathematical tables and formulae*, *30*, 569–668.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, *21*(9), 649–665.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive psychology*, *104*, 57–82.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533–558.