Communicating understanding of physical dynamics in natural language

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

Jane Yang
Department of Cognitive Science
UC San Diego
j7yang@ucsd.edu

Ronen Tamari
School of Computer Science
Hebrew University of Jerusalem
ronent@cs.huji.ac.il

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

Abstract

Our ability to share abstract knowledge with others is a defining feature of modern human intelligence. What information do people choose to include in such communication? Here we develop a novel physics-based video game to elicit natural language responses on how this game works to teach other people. We collected data from 238 participants and found that people explicitly described the latent physical properties of the game environment like mass and gravity in their responses. We also found that people who performed better in the game also produced responses that covered more latent physical properties. Taken together, our study provides novel insight into how people communicate their understanding of physical dynamics in natural language.

Keywords: intuitive physics; explanation; communication; linguistic abstractions

Introduction

Much of what we learn about the world comes not from our own experience but by learning from others, often via language. For example, coaches often instruct kids on how different strokes affect the trajectory of a tennis ball; we were taught since little not to sit in the way of wind to prevent inhaling smoke at a barbecue. Our ability to transmit and build upon abstract causal knowledge previously learned by others is a fundamental aspect of modern human cognition (Tomasello, Kruger, & Ratner, 1993; Boyd, Richerson, & Henrich, 2011). This propensity for sharing abstract and generalizable knowledge has enabled us to accumulate rich information about the structure of our world. How then do people communicate about their structured knowledge to help others to learn? What information do they choose to include?

Prior work investigating communication suggests that explanations provide an important means of conveying abstract causal knowledge (Lombrozo, 2006), yielding benefits for one's own (Chi, De Leeuw, Chiu, & LaVancher, 1994) or others' learning (Webb, 1982). Explanations often go beyond what can be experienced, conveying abstractions and generalizations (Voiklis & Corter, 2012; Schwartz, 1995). In particular, it has been shown that *generic language* (e.g. "Birds fly.") provides a simple and ubiquitous way to communicate generalizations about categories (Tessler & Goodman, 2019). It has also been studied how people communicate about the causal relationship between different parts in a mechanical systems (Huey, Walker, & Fan, 2021). However, our rich knowledge about the world goes beyond just knowing about concepts and detecting the mere presence of causal relationship between variables: people build structured generative *world models* that provide reliable predictions so as to behave appropriately in our open-ended, dynamic environment and generalize to various scenarios (Battaglia, Hamrick, & Tenenbaum, 2013). Thus, how people are able to communicate about their world model encoding the dynamics of the physical environment poses a different challenge than communicating about concepts or causal relationships. A more detailed characterization of explanations on physical world models is critical for advancing our understanding of how people transform their everyday direct experience with the physical world into compressed representations that explain how things work.

A wealth of research in intuitive physics has established that people have rich world models that encode the underlying dynamics of the physical world (Sanborn, Mansinghka, & Griffiths, 2013; Ullman, Spelke, Battaglia, & Tenenbaum, 2017; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018). Such models not only enable people to make reliable physical predictions and inferences but also generalize to unpredictable future tasks and situations, regardless of their formal understanding of physics. However, it is less clear how people are able to convert this *implicit* understanding of the physical world to *explicit* knowledge that can be transmitted to others. A better understanding of how people communicate about their intuitive physics knowledge will not only shed light on how people encode physical abstractions in language, but also supports a broader conceptualization of how people's internal world model is structured.

In this paper, we investigate how people transform their direct intuitive physics experience into compressed forms to teach other people using language, and identify what information people choose to include in such communication. To this end, we developed a novel physics-based video game to elicit natural language explanations on how this game works. Specifically, participants were asked to use a paddle to catch balls of different masses in two environments where different latent forces (e.g. gravity, wind) are at play. We then ask them to write a short paragraph to teach someone who has never played this game how this game works. To identify the distinctive information people choose to include, we use a baseline group for comparison where responses were produced in the absence of any explicit goal to teach an audience how to play the game. Our results suggest that people

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

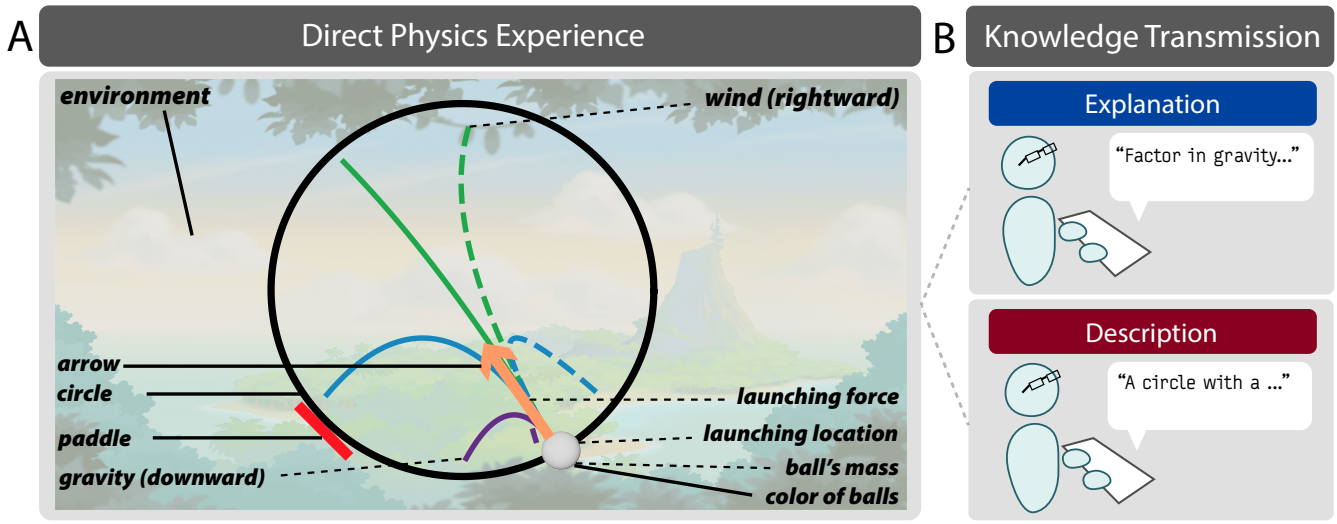

Figure 1: Two-stage experiment. (A) Stage one: direct physics experience through the game, participants played a physics-based video game to catch the ball with the red paddle under different settings. Physical concepts are annotated in dashed lines, visual concepts in solid lines. (B) Stage two: natural language elicitation, participants were randomly assigned to either explain or to describe the game upon finishing the game.

can successfully perform this task and that the latent physical properties like mass and gravity were explicitly encoded in their language. Furthermore, we found that those who performed better in the game also mentioned more latent physical concepts.

Experiment

Participants

238 participants (142 female; $Mean_{age}$ = 31.8 years) recruited from Prolific completed the experiment. We included data from all participants, as they all met our preregistered inclusion criteria. Participants provided informed consent in accordance with the UC San Diego IRB. The experiment lasted approximately 40 minutes and participants were paid $14/hr based on this expected completion time.

Task environment

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 (see Figure 1A for an example). 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 happy with the paddle's location, they launched the ball using the spacebar and 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 preform this task, the mass of the ball, the location where the ball was launched, and the force the ball was launched with. Each of these design elements is explained below.

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 2A). Each ball-environment combination consists of 24 trials (12 launching locations $\times$ 2 launching forces), resulting in 144 trials in total.

In order to probe generalization and thus deeper understanding of physical dynamics, we divide the game into a training phase and test phase, the division was not visible to participants. In the training phase, participants are exposed to five of the six ball-environment combinations. The subse-

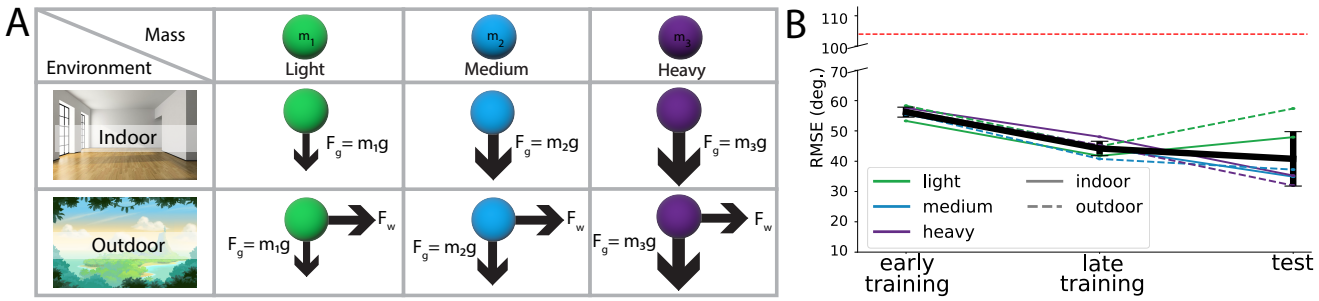

Figure 2: (A) 2×3 design matrix, where participants were trained on 5 out of these 6 cells and asked to generalize to the other cell. (B) RMSE for all 6 conditions, black thick line is the mean. The dashed red line represents expected performance under random guessing (Rinaman et al., 1996).

quent test phase only includes trials with the remaining ball-environment combination. In order 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 the six ball-environment combinations as their held-out test phase.

Natural language elicitation

After the participants have finished the game, they were randomly assigned to either explain or to describe (see Figure 1B), resulting in 119 participants in each condition. In the explanation condition, they were given the following prompt:

"*Imagine someone who has never played this game before but is interested in learning how to play. What would you tell them so that they understand where to put the paddle on any given trial? Please write a short paragraph that provides them with the most important information to help them succeed.*"

in the description condition, they were prompted to:

"*Imagine someone who has never played this game before but is interested in picking it up at the store. What would you tell them so that they could identify it based on a few screenshots? Please write a short paragraph that provides them with the most important information to help them pick out this game from a lineup of other similar-looking ones.*"

Results

How well can people play this game?

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 2B shows RMSE for all 6 conditions. Participants' performance was significantly above chance at every point during this experiment ($t = -73.16$, $p < 0.001$). Initially, RMSE was high (mean=56.22°), 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 1A 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=40.79°; $b = -6.55, t = -2.90$, $p < 0.01$). 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. Broadly, this suggests that while people may have struggled to learn the mechanics of the task at the beginning, they rapidly improve over time.

How are explanations different from descriptions?

Given participants' successful learning and generalization, the next question we ask is how were people able to convey their direct physical experience in natural language such that they can explain how this game works to someone else. In other words, what information do people choose to include in their explanations? One hypothesis is that people will choose to directly convey their raw experience. Critically, on this view, they will put an emphasis on the *observable* visual features of the game. However, another possibility is that participants will emphasize the mechanics of the game but ground it in the visual attributes. In this view, rather than only conveying information about the *observable* features of the game, people will distill their raw experience into the *latent* dynamics of the physical world that they inferred from raw experience, but ground it in the visual components of the game to better explain to a novice about how this game works. Finally, a third account is that explanations fully abstract away from

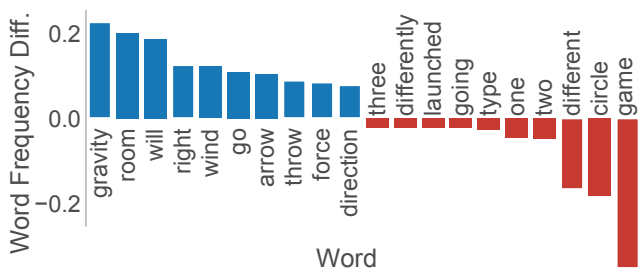

Figure 3: Difference of normalized word frequency between explanations and descriptions. Blue bars represent words that are more frequently used in explanation than description, red for the other way around.

the surface details, and just preserve the functional information, omitting mentioning perceptual features.

To test these hypotheses and identify the properties that are distinctive of explanations, we use description as a baseline for comparison, which were produced in the absence of the goal to teach someone how to play this game. In this section, we first analyze the raw responses from the two conditions. We found that explanations produced by participants tended to be longer than the produced descriptions in terms of word counts ($Mean_{exp} = 63.14$, $Mean_{des} = 41.12$; $b = 22.03$, $t = 4.54$, $p < 0.001$), suggesting that people devoted more effort and contained more information when producing explanations than descriptions. We further observed that the distribution of word count for explanation is long-tailed, meaning that some participants chose to write long paragraphs when prompted to explain, whereas the word count distribution for the description group is more concentrated at the mean. To rigorously evaluate whether these two distributions are different, we computed the Jensen-Shannon distance (JSD) between the word count distributions of the two conditions. We compared the JSD to a null distribution generated by randomly assigning responses to the two conditions (Nichols & Holmes, 2002) and found that the distance between the two conditions was greater than that expected under the null ($p < 0.05$).

Beyond word counts, to explore any systematic differences in the words used in explanations or descriptions, we investigated the frequency of each word used in both conditions. We visualize the difference of the normalized frequencies in Figure 3. This demonstrates that a variety of concepts from both latent physical properties and visual components were presented in language responses collected. Nevertheless, words pertaining to game mechanics and latent physical properties ("wind", "gravity", "force", "direction") were more frequently mentioned in explanations, whereas words that describes the game ("circle", "one", "two", "three") were more frequently used in descriptions.

The overall statistics in word counts and most diagnostic words for each condition provides us with insight on the *literal* overlap in participants' responses when prompted to either explain or describe. We next sought to explore how the information being conveyed in the two conditions are *semantically* different. To explore this question, we leveraged the rich semantic information encoded in the large pre-trained neural language models developed recently by extracting the embeddings of these responses using the language model. Specifically, we used the Sentence Transformers (Reimers & Gurevych, 2019) to obtain a contextual vector embedding (dimensionality $d = 384$) for each response and then applied dimensionality reduction to project the response embeddings to a 2D space for visualization (see Figure 4).

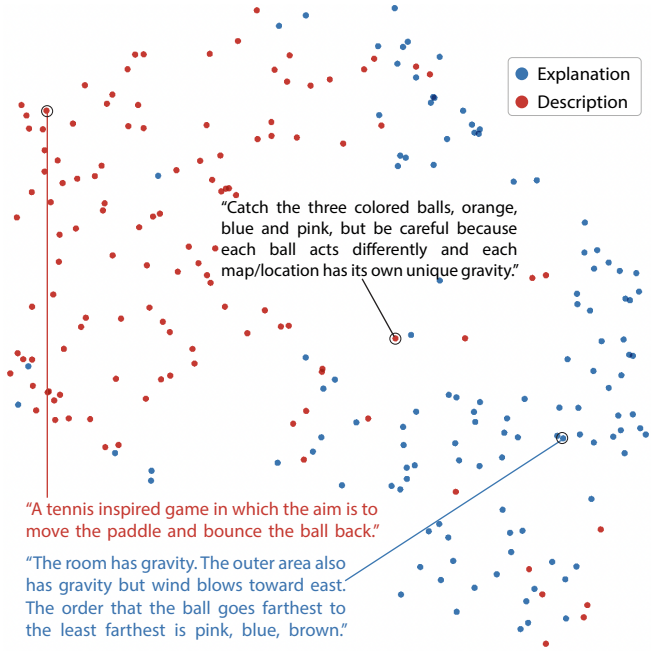

Figure 4: 2D projection for neural language model vector embeddings of responses, showing high degree of separation between explanation and description responses. Three data points were annotated as examples for a typical explanation response, a typical description response, and a boundary response, respectively.

For a simple measure of the separability of the two response types (Lorena, Garcia, Lehmann, Souto, & Ho, 2019), we fit a linear support vector machine classifier (Cristianini, Shawe-Taylor, et al., 2000) to the embeddings. We use 5-fold cross-validation (0.7/0.3 train/test ratio) and obtain a mean accuracy of 0.92 ± 0.01. This relatively high degree of separation indicates significant *semantic* differences between the language used by participants in these two conditions (see the three annotated responses as typical examples for explanation or description in Figure 4). Together, these findings suggest that participants used different language both in terms of linguistics and semantics when prompted to explain or describe.

How well do the concepts people mentioned correspond to the task environment?

In the previous section, by analyzing raw linguistic data, we found that responses from explanation and description are significantly different form each other in terms of word counts, most diagnostic vocabularies, and latent embeddings,

but this does not provide us with a clear insight on how the information being conveyed in the two conditions are *grounded* in the specific relevant concepts for this task environment. To this end, we annotated the collected linguistic corpus using 13 different tags. We focused on two domains of conceptual content used in this game. First, we have 6 tags for language that conveys the physical properties in the game: these are the physical parameters that define a generative model where each trial is sampled from, thus determine the trajectory of the ball, and where the participants should place the paddle. These 6 tags covers the main components of this generative model. Given participants' successful learning and generalization, we hypothesize that they should be able to cover the basic components of this generative model in their language response when prompted to give a good explanation for how to play this game. Second, each of the latent physical parameters only becomes observable by being bound to visible attributes of each scene – these are what participants directly observe from on a single trial. To accommodate this, we made a tag for every visual component on the task display, resulting in another 6 tags. The last tag was for miscellaneous information (see Figure 5 for all the tags we used). The annotation of language responses using tags was performed by one of the authors blind to the prompt condition (i.e. explanation/description).

Leveraging these semantic part annotations, we found that the distribution of tags are significantly different across the two conditions ($\chi^2(12) = 251.98$, $p < 0.001$): explanations invoke the key physical concepts much more frequently than descriptions, highlighting the centrality of these concepts for explanatory responses. Both descriptions and explanations invoke visual concepts, though these are more common in descriptions. The reduced emphasis on visual concepts in explanations replicated the findings in the literature on verbal explanations (Legare & Lombrozo, 2014) and on visual explanations (Huey et al., 2021), in which explanations had a reduced emphasis on perceptually salient but irrelevant features.

It is worth noting that even though explanations had a systematic emphasis on physical concepts compared to descriptions, a number of visual concepts (e.g. color of the balls, background image, etc.) were still frequently mentioned. This is because physical concepts were grounded by the game environment, participants often need to utilize *visual* concepts to establish a working context for *physical* concepts they want to explain, in other words, physical concepts would not be sensible without the presence of visual features. For example, in "*View the three balls as being one heavy (brown), one medium (pink), one light (blue). The room has normal gravity and the forest has offset gravity to the bottom right*", the visual concepts, "brown", "pink" (color of the balls), "room", "forest" (background image), etc., provided a critical context to explain the concept of mass and forces. Together, these findings suggest that when prompted to explain, people neither only focused on visual attributes of the game, nor did they fully abstracted away from the surface details, but rather explained the latent physical properties and mechanics by grounding them in the visual components.

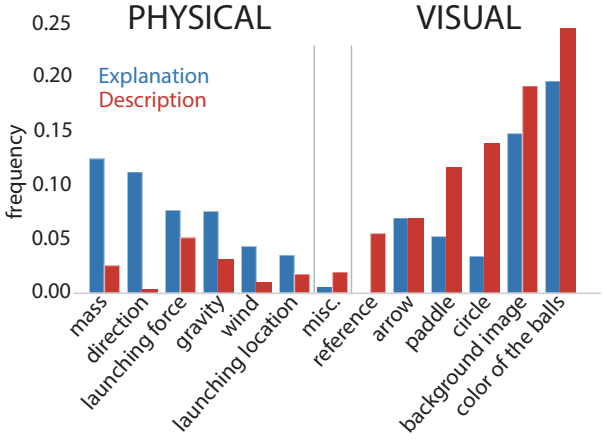

Figure 5: Frequency of different tags for the two conditions, normalized by the total number of tags in each condition.

Do all explanations put a strong emphasis on the latent mechanics of the game? We next sought to explore the variability of responses *within* the explanation condition. To this end, we designed another way of tagging the responses: explanations were broken down into two categories, causal or procedural. A causal response encodes the *latent* dynamics of the game but does not specify how someone should exactly place their paddle on a given trial. A procedural response, on the other hand, details a *procedure* that someone should follow, but does not identify any underlying causal forces or relations. See below for an example of each category:

Causal: "*The blue ball is heavy and immediately drops close to where it lands, the yellow is light and is less impacted by gravity, the red is in the middle point of these. In the room there is no wind effecting the balls path but outside there is a wind blowing the balls to the right to factor into the balls flight path.*"

Procedural: "*...for the blue ball, you should often place the paddle across the circle from where the ball launches, although it has a tendency to move to the left. For the pink ball, it does not launch very far and tends to fall. If the pink ball is launching from the bottom, place the paddle slightly to the left of the ball because it will fall and will not move far. If the pink ball is launching from the top of the circle, place...*"

In our initial exploration of the prevalence of "causal" and "procedural" explanations, one of the authors annotated each response. This preliminary analysis suggests that causal explanations (58.14%) were more prevalent than procedural explanations (23.26%), with some explanations not falling neatly into either category (18.60%). To provide some external validation that the explanations that had been tagged as "causal" and "procedural" were semantically distinguishable, we conducted 5-fold SVM classification analysis on their Sentence Transformer embeddings, and found that they could be classified with reasonably high accuracy (0.88 ± 0.07).

How does the content of explanations relate to task performance?

Previous research has suggested that when instructing novices, experts will make more abstract statements and fewer concrete statements than beginners (Hinds, Patterson, & Pfeffer, 2001; Chi, Feltovich, & Glaser, 1981), as well as use more advanced concepts and fewer basic concepts (Hinds et al., 2001). In line with this idea, we hypothesize that participants who performed well on this game should be able to provide more abstract explanations when prompted, and at the same time mention more latent physical concepts in their responses than those who performed less well in the game.

To this end, we define a simple metric to quantify the quality of a response with respect to the two kinds of semantic concepts discussed above: physical and visual. We define R_P and R_V as the recall over the set of physical concepts P and visual concepts V, respectively. Denoting $T(x)$ to be the set of tags appearing in response x, we can calculate physical/visual recall for a given response x as follows:

$$R_P(x) = \frac{|T(x) \cap P|}{|P|}, R_V(x) = \frac{|T(x) \cap V|}{|V|}$$

A response with perfect recall for physical and/or visual concepts means that all relevant corresponding concepts were covered. We use participants' test phase performance as a proxy for task performance, which was measured by their RMSE compared to the ground truth landing locations (lower RMSE means better performance).

We group responses by visual and physical recall respectively. Responses with above median physical recall R_p exhibit a lower RMSE compared with responses with below-median R_p (above: $32.70°$, below: $43.06°$; $t(116) = 3.27$, $p < 0.01$). For the case of visual concepts, we did not find a significant difference on performance between groups (above: $35.18°$, below: $39.18°$; $t(116) = 1.11$, $p = 0.27$). Furthermore, we found a negative correlation between physical recall R_p and RMSE ($r = -0.32$, $p < 0.001$), but did not find a strong correlation between visual recall R_v and RMSE ($r = -0.18$, $p = 0.06$). Taken together, these results suggest that participants who performed well on this game tend to mention more latent physical concepts compared to those who performed less well, echoing previous findings which suggests that experts use more advanced concepts and fewer basic concepts.

To test our hypothesis about the interaction between performance and abstractness, we compare the RMSE of the causal explanations and the procedural explanations, and found a significant difference (causal: $33.96°$, procedural: $43.98°$; $t(116) = 3.01$, $p < 0.01$). We further compared the percentage of each type of explanation in different performance groups. As is shown in Figure 6, the distribution of explanation types for participants performing above the median is different from that for below-median participants ($\chi^2(2) = 10.49$, $p < 0.01$). Among participants performing above the median on the task, the proportion of causal explanations is much higher that those in the bottom half. This suggests that participants who performed well on this game indeed tend to give more abstract causal explanations as opposed to providing detailed procedures.

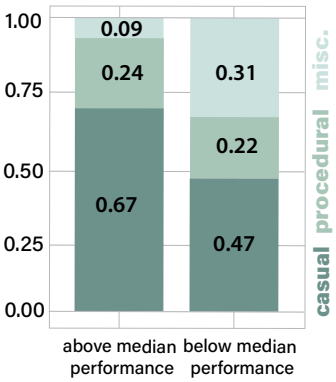

Figure 6: Proportions of causal and procedural explanations in different performance groups. Category misc. is for explanations that do not fall neatly into either category, e.g., "*The game is great but slightly missing the ball is not fun. Especially after thinking you have mastered it*".

Discussion

How do people transmit knowledge about intuitive physics to others? This paper investigated how direct intuitive physics experience was compressed into language and identified the critical information people chose to include in such communication. We developed a novel online physics-based video game to simulate intuitive physics experience and asked participants to teach someone else to play this game in natural language. By comparing to a baseline group, we found that people chose to explicitly describe the *latent* physical properties of the game environment like mass and gravity when teaching others. We also found that those who performed well in this game provided more abstract explanations and at the same time covered more latent physical concepts rather than concrete visual features in their response than those who performed less well in the game.

Knowledge built culturally across generations allows humans to learn far more than an individual could glean from their own experience in a lifetime. Our work sheds light on how abstract causal knowledge about intuitive physics can be transmitted between individuals. In future work, we plan to investigate how people learn to play intuitive physics games *guided* by natural language explanations, possibly more efficiently. Future experiments could assign each participant as either *teacher* or *student* where the teacher plays the game first and pass along explanations to the student, who proceeds with the explanation to explore the underlying dynamics of the game. Investigating the student's learning efficiency and generalization behavior will be critical to expose the role of culturally transmitted knowledge in individual's learning.

In sum, our paper reveals novel insights about how intuitive physics experience was compressed into natural language in order to teach someone else. Insights from such studies may lead to a deeper understanding of how we encode and explain abstract physical knowledge as well as facilitating AI research in human-machine collaboration.

Acknowledgments

We would like to thank 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.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*(Supplement 2), 10918–10925.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, *18*(3), 439–477.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, *5*(2), 121–152.

Cristianini, N., Shawe-Taylor, J., et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Hinds, P. J., Patterson, M., & Pfeffer, J. (2001). Bothered by abstraction: The effect of expertise on knowledge transfer and subsequent novice performance. *Journal of applied psychology*, *86*(6), 1232.

Huey, H., Walker, C., & Fan, J. (2021). How do the semantic properties of visual explanations guide causal inference?

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of experimental child psychology*, *126*, 198–212.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464–470.

Lorena, A. C., Garcia, L. P., Lehmann, J., Souto, M. C., & Ho, T. K. (2019). How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, *52*(5), 1–34.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, *15*(1), 1–25.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. No. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *The journal of the learning sciences*, *4*(3), 321–354.

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological review*, *126*(3), 395.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and brain sciences*, *16*(3), 495–511.

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.

Voiklis, J., & Corter, J. E. (2012). Conventional wisdom: Negotiating conventions of reference enhances category learning. *Cognitive Science*, *36*(4), 607–634.

Webb, N. M. (1982). Student interaction and learning in small groups. *Review of Educational research*, *52*(3), 421–445.