UNIVERSITY OF CALIFORNIA SAN DIEGO

Learning structured representations for generalization in the physical world

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Experimental Psychology

by

Haoliang Wang

Committee in charge:

Professor Timothy Brady, Co-Chair
Professor Judith Fan, Co-Chair
Professor Nadezhda Polikarpova
Professor Caren Walker

2024

The Dissertation of Haoliang Wang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

EPIGRAPH

"[The brain is] perpetually weaving a picture of the external world, tearing down and reweaving, inventing other worlds, creating a miniature universe."
Wilson (1999)

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Judy Fan. Thank you for showing me how to do science from day one of grad school, and always challenging me to think deeper and look farther. For your patience and tolerance with my naivety and often half-baked ideas. For being so generous with your time to discuss topics outside research: philosophy, movies, science fictions, career and life. For helping me become a better scientist and also a better person. Thank you for being a great mentor.

I would like to thank Nadia Polikarpova. Nadia welcomed me to her group at the very beginning of my grad school when I reached out for collaboration. Thank you for helping me navigate through a different field and showing me the ropes. For sharing opportunities and connecting me to others.

Thank you to other members of my committee: Tim Brady and Caren Walker for providing insightful feedback from different angles on earlier versions of this thesis.

I would also like to thank Ed Vul, for your participation in the first half of my grad school. I have always been amazed by your wit in computational methods and your sharpness in writing precise and concise arguments. And also thank you for the humor you brought to our meetings.

Thank you to all of my co-authors: Kevin Smith, Kelsey Allen, Josh Tenenbaum, Rahul Venkatesh, Khaled Jedoui, Dan Yamins, Robert Hawkins and many others. The conversations and collaborations I had with you all increased my sense of belonging to the community, and made this Ph.D. journey much more fun.

I would like to thank the amazing members of the cogtools lab: Will Mccarthy, Holly Huey, Erik Brockbank, Felix Binder, Charles Lu and others, for creating such a fun environment both in and outside the lab. I appreciate your always constructive feedback on various presentations I gave in the lab. I also would like to thank my wonderful research assistants: Jane Yang and Nora Chen. I cannot wait to see you begin your new journey.

I am grateful for the all the friends I met in grad school, especially: Thank you to

Yang Wang and Wenhao Qi, for your company throughout grad school and also for being a foodie with me and exploring all the restaurants and boba tea in San Diego. Thank you to Shuai Shao, I couldn't have asked for a more supportive and understanding office mate.

Finally, thank you to my parents, for always trusting and supporting my choices, even when they were risky.

Chapter 1, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Wang, H., Polikarpova, N., & Fan, J. (2021). Learning part-based abstractions for visual object concepts. In *Proceedings of the 43th Annual Meeting of the Cognitive Science Society*. The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Wang, H., Allen, K., Vul, E., & Fan, J. (2022). Generalizing physical prediction by composing forces and objects. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Wang, H., Jedoui, K., Venkatesh, R., Binder, F., Tenenbaum, J., Fan, J., Yamins, D., & Smith, K. (2024). Probabilistic simulation supports generalizable intuitive physics. In *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*. The dissertation author was the primary investigator and author of this material.

VITA

2019	Bachelor of Engineering, Automation, Xi'an Jiaotong University
2022	Master of Arts, Experimental Psychology, University of California San Diego
2024	Doctor of Philosophy, Experimental Psychology, University of California San Diego

ABSTRACT OF THE DISSERTATION

Learning structured representations for generalization in the physical world

by

Haoliang Wang

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2024

Professor Timothy Brady, Co-Chair
Professor Judith Fan, Co-Chair

Humans must continually generalize from past experience to novel scenarios. For instance, when we see a new coffee mug, although it does not look exactly the same as the ones we have seen before, we can still categorize it as a coffee mug, recognize the handle, and infer that it will probably break when dropped to the ground. What mental representations enable people to make predictions about objects in new environments? In this dissertation, I explore different possible computational constraints on what people infer about object properties and their interactions that enable them to generalize. In Chapter 1, I begin by investigating how people learn to recognize novel objects. I develop a paradigm where I systematically manipulate the underlying rules that generated the stimuli. I find that people can learn both specific exemplars and abstract object patterns depending on the statistics of the environment and generalize to novel instances. I show that people's behavior can be explained by a program learning model that simulates the generating process of the stimuli. In Chapter 2, I analyze how people make generalizable judgements about the physical interactions between objects. People are asked to predict the behavior of an object in an environment it has never been in before. Results suggest that people simultaneously infer multiple physical variables (i.e., object mass, force in the environment) based on past observations of how objects behave and compose knowledge about these variables to make predictions. In light of these findings, Chapter 3 examines how people generalize their understanding of physical interactions across multiple scenarios. Across two studies with seven different physical scenarios (e.g. collision, containment, support, etc.), I find that people are accurate and consistent with each other in their predictions about whether and where two objects would contact. I then show that a computational model that runs simulations of noisy physics generalizes in human-like ways across all scenarios. Altogether, this dissertation suggests that representing entities and how these entities interact in an abstract internal model plays a key role in, and potentially provides a coherent account of, human generalization.

Introduction

0.1 Overview

Life rarely, if ever, replays exactly the same events, with exactly the same objects that we have experienced before. We are constantly faced with the problem of generalizing from past experience to novel scenarios, and it would be hard to imagine what our lives would be like without such an ability: we might know the things we had experienced in the past – a particular coffee mug, for example – but when we go into a new room and see a new coffee mug, we would be at a loss. We would have to study it from scratch, attempt to determine whether it is alive or dead, what its function is, whether it will hurt us, or whether it will float in the air if pushed off the table. What are the underlying representations and computations that enable us to robustly generalize? The work in this dissertation explores the possibility of building a unifying computational framework for understanding the mental representations of human generalization across different domains. At the heart of this framework is the hypothesis that people are not just learning to map input sensory information to output actions, but are rather inferring the latent properties and structure of their environment, forming an internal "world model", which then enables them to extend to novel objects and situations. To explore this hypothesis, my research combines insights from psychology and computer science to reveal how this abstract mental model of the world can support both human visual perception and physical prediction. The results point the way towards building a unifying computational model of how humans generalize to novel situations.

0.2 The challenge of generalization

"Because any object or situation experienced by an individual is unlikely to recur in exactly the same form and context, psychology's first general law should, I suggest, be a law of generalization". In his paper "Toward a Universal Law of Generalization for Psychological Science (Shepard, 1987)", Shepard argued for the importance of studying the behavior of generalization in psychology. Indeed, looking around ourselves, the need for generalization cannot be overemphasized: From recognizing familiar faces in different settings to navigating novel situations, we need to routinely generalize our knowledge and behaviors in order to adapt to the complexities of our environment. Over the last several decades, the question of generalization has received extensive theoretical and empirical research interest in psychology for its fundamental role in cognition (Harnad, 2017). The earliest work on this topic dates back to the beginning of psychological research itself: Pavlov's classic experiments on classical conditioning (Pavlov, 2010) demonstrated how animals generalize learned responses to stimuli that are similar to the conditioned stimulus. Pavlov's work laid the groundwork for understanding the principles of learning and generalization and since then, generalization has been studied empirically in almost every sub field of psychology: from how people generalize vocabulary from specific instances to broader linguistic contexts (Goldberg, 2009; Richtsmeier et al., 2011), to how positive and negative experiences guide future goal pursuit and emotional reactions (Lerner & Keltner, 2000); from how individuals generalize past experiences of reward and punishment to make decisions in novel situations (Gonzalez & Dutt, 2011; Hau et al., 2008; Tversky & Kahneman, 1974), to how a group of people coordinate and form new conventions in social settings based on past interactions (Hawkins et al., 2023; Lewis, 2008; McCarthy et al., 2021).

The work in this dissertation focuses on people's generalization behavior in the visual world: Perhaps the first and foremost challenge faced by any agent in this world is to establish a generalizable understanding of the objects that exist in the world (*visual concept learning*), as well as how they interact with each other (*intuitive physics*). This is a prerequisite for any kind of more sophisticated generalization in humans. However, it is never an easy problem. One of the primary difficulties of generalization in the visual world stems from the variability of stimuli: The world is replete with diverse objects, events, and situations that have multiple features and characteristics. Any agent in this world can only experience a very small fraction of it, how do we generalize to new environment from such sparse data?

At first glance, this challenge seems daunting. Indeed, there is a wealth of evidence demonstrating different kinds of errors people make in both recognizing objects and predicting how they interact with each other in the face of novel situations. For example, research in developmental psychology suggests that children tend to "overgeneralize" in categorization early in development (Ambridge et al., 2013; Saltz & Sigel, 1967), the idea being that when children are only exposed to a limited set of instances within a category, they may become overly tuned to the specific features of those examples, thereby impeding their ability to generalize their knowledge to wider situations or instances. To illustrate, imagine a child learning to identify dogs. If they only ever see pictures of golden retrievers during their training, they might become overfitted to that particular breed and have difficulty recognizing other breeds or even different types of animals as dogs. Similarly, the complexity of the world also poses challenges for generalizing in physical reasoning, and this is the way the field of intuitive physics started — by noticing the apparent errors people make when presented with various physical prediction problems. Mccloskey and colleagues found that a large proportion of participants believed that a ball will continue to move in a curved path after leaving a curved tube, although the reality is that it will travel in a straight line according to Newton's laws (McCloskey et al., 1980; McCloskey & Kohl, 1983). Subsequent research in intuitive physics identified more errors people make when generalizing to different scenarios, for example, Caramazza et al. (1981) and Hecht and Bertamini (2000) showed that people lacked a basic understanding of projectile motion even though it is ubiquitous in life; McCloskey et al. (1983) showed that participants tended to believe that things would fall straight down after being released from a moving carrier; Hecht and Proffitt (1995) and Howard (1978) showed that people also hold incorrect beliefs about nonrigid bodies: When shown a tilted container, people often do not realize that the surface of the liquid in the container should remain horizontal with respect to the ground.

In spite of the many failures from people when extending to novel situations, there has also been evidence demonstrating people's remarkable proficiency in forming accurate and useful generalizations. This has been studied in both visual concept learning and intuitive physics. For instance, Lake et al. (2015) showed that people could learn a novel visual concept from just one example and generalize to both recognizing and generating new instances of that concept; furthermore, Xu and Tenenbaum (2007) showed that people could not only learn new concepts from limited evidence, but they could also infer the extension of the concept (e.g. on which level of the semantic hierarchy that concept is on). Similar findings were observed in intuitive physics: Hamrick et al. (2016) showed that participants had the capacity to reason about the relative masses of different objects after watching a video of a tower falling or not falling in naturalistic 3D scenes; Ullman et al. (2018) extended this to inferences on more physical properties of objects (e.g. friction) and environment (e.g. global force) from observing a short video clip of objects interacting in a 2D world.

What explains people's errors and what accounts for their successes when generalizing to new experiences? Implicit in Shepard's proposal about generalization (Shepard, 1987) is the notion that stimuli are embedded in an abstract "psychological space" in people's minds – a *representation* of past experience that facilitates generalization in the future. In what follows, I review prior **theories** of generalization that make different assumptions about the underlying representation. I describe questions not answered by this prior research and how this dissertation builds on and contributes to the literature. Finally, I present a brief outline of the results from my work and how this may lead towards a more comprehensive account of how we generalize our understandings of objects and their interactions to novel situations.

0.3 Existing theories of generalization

All theories of generalization are about abstraction. Proper generalization requires forming the right abstraction of past experiences over which similarities can be assessed between the past and the present (Son et al., 2008). What is different between theories is the assumption about *how abstractly* past experiences is represented. Here, I review theories proposed in the visual concept learning and intuitive physics literature that aim at answering the question of how people generalize. As we will see, these theories lie on a ladder: a spectrum of how much abstraction is involved when encoding past experiences, from more concrete to more abstract (see Figure 1).

0.3.1 Exemplar-based approaches

At one extreme of this spectrum, it is hypothesized that there is minimal abstraction when encoding the past: we store the raw or nearly raw experiences into memory and generalization is a process of retrieving instances from memory and assessing the similarity between the instances retrieved and the novel instance being presented. This hypothesis has many instantiations across different domains but perhaps the most representative one is the exemplar theory in categorization (Medin & Schaffer, 1978; Nosofsky, 1986, 1984; Shi et al., 2010). Exemplar theory suggests that people categorize objects and concepts by comparing them to specific examples (hence the name, exemplar theory) stored in memory. Each category is represented by the set of all remembered exemplars of that category. When people encounter a new stimulus, they compare it to the stored exemplars to determine its category based on their similarity. The category to which a new stimulus is assigned depends on the cumulative similarity between the stimulus and all the stored exemplars of various categories. The category with the highest summed similarity is chosen. The exemplar theory has clear advantages as a theory of generalization in concept learning. For example, it can explain the flexibility of human categorization as it does not rely on

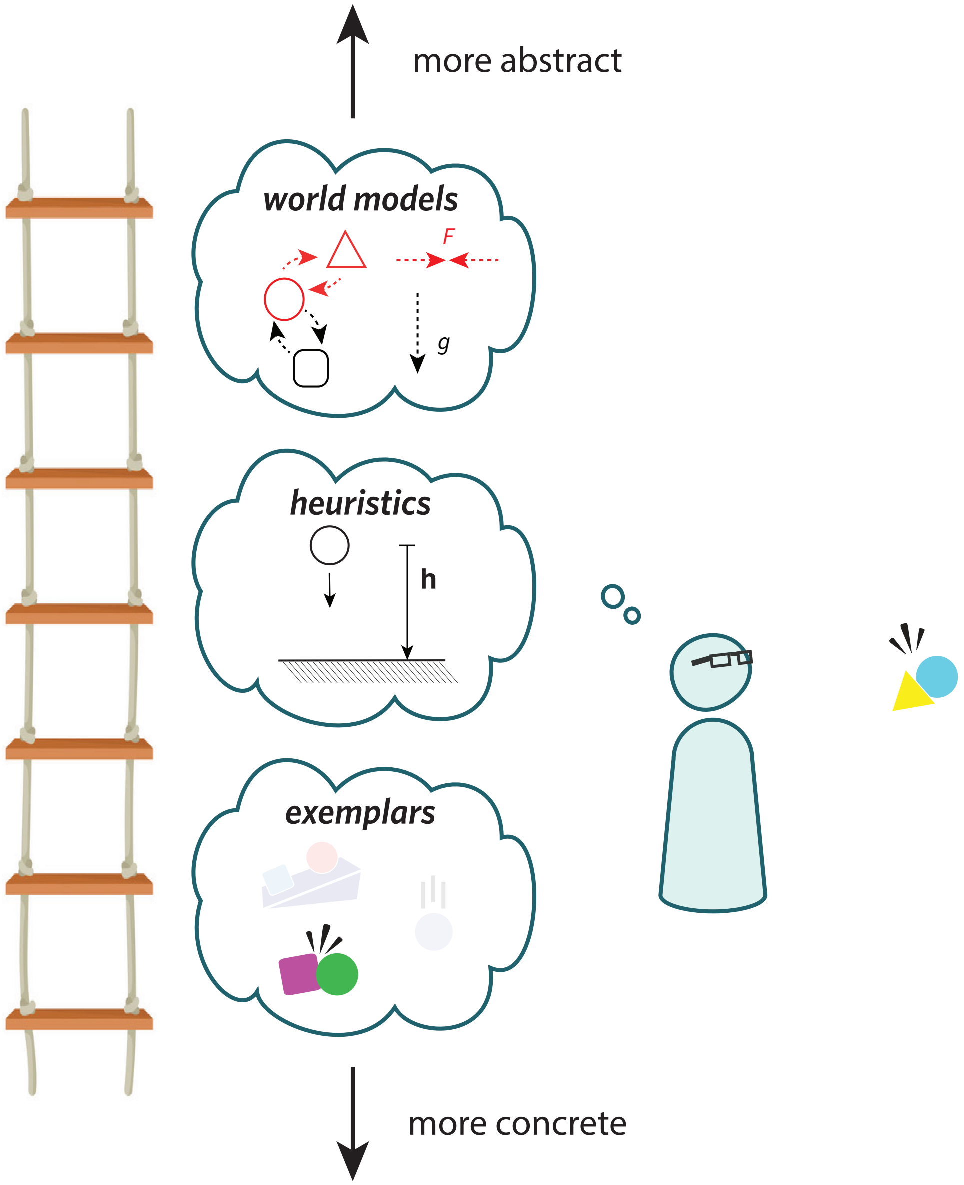

Figure 1. Theories of generalization differ in assumptions about how abstractly past experiences is represented, different theories lie on different levels of the "ladder of abstraction". At the bottom level of the ladder is exemplar-based approaches, where concrete experience is stored into memory and compared against the novel situation when generalizing. Heuristics is one level up on the ladder, assuming summary statistics or particular features extracted past experiences should suffice when generalizing to new environments. On the top level of the ladder is the world model hypothesis, which posits that people have an internal causal model of the external world incorporating knowledge about objects, events, and the relationships between them.

an average representation but rather on multiple specific instances. As a result, it can also explain why people might categorize borderline or atypical instances in a particular way based on past experiences. However, there are also obvious limitations of this hypothesis. The most significant one concerns cognitive load and memory capacity: Storing a large number of specific exemplars for each category can be demanding on memory, and even if large number of exemplars can be stored, comparing new stimuli to each exemplar can be cognitively very demanding.

Arguments that are similar to the exemplar theory have been made in intuitive physics as well. Across two experiments, Kaiser et al. (1986) asked participants to predict trajectories of objects exiting a curved tube (similar to McCloskey et al. (1980)). The authors found that people were more accurate on the familiar version of the problem compared to a novel, abstract problem. The authors further argued that the way participants solve the physical reasoning problems is to draw on specific experiences, and that people are able to reason more appropriately about motion problems when they are related to specific, concrete, familiar experiences; they performed worse on the abstract problem because it failed to evoke any specific memory.

0.3.2 Heuristics

On the ladder, another class of theories that assumes less specificity in encoding experiences is heuristics (one level up in the abstraction ladder, see Figure 1). Compared to the exemplar-based approaches where concrete experience is stored and used for generalization, in heuristics models past experiences is abstracted as mental *shortcuts*. For instance, one heuristics when judging the stability of everyday objects would be "if something is stacked too high then it will fall over". This is true in most cases that we have seen, so it is *likely* that a new Lego tower will also fall over if it has too many layers. Just as this example illustrates, heuristics allow people to simplify complex problems by focusing on only a few relevant factors that are proved to be useful in the past (the height of the tower in this example), as opposed to storing concrete instances or situations — a nice summary statistics of the past abstracting away the details. Heuristics have been a prominent account of how people perform physical reasoning: When making physical predictions, people might base their judgments exclusively on combinations of features of the initial scene configuration (Gilden & Proffitt, 1989; Kozhevnikov & Hegarty, 2001; Nusseck et al., 2007; Proffitt et al., 1990; Sanborn et al., 2013; Todd & Warren Jr, 1982). For example, Paulun et al. (2023) showed that in the case of asking people to judge the elasticity of bouncing objects, the duration of the bouncing movement gave a decent estimation of the property and captured participants responses reasonably well. Heuristics make up for the deficiencies of exemplar-based approaches as they do not require assessing the similarity between the novel situation and each of the stored exemplars, rather, applying simple rules of thumb allows people to make fast and efficient judgements. Nevertheless, a major drawback of heuristic models is the possibility of cognitive biases. These biases stem from the simplifications inherent in heuristics, which often neglect the underlying causal mechanisms of past experiences during encoding. Using the stability heuristics example mentioned above, this heuristics does not always lead to correct predictions. For example, consider a block tower where the objects are glued together, obviously it is far less likely to fall over, but the simple heuristics would still make the same prediction as it would make for any other tower because it does not reference to the underlying physical dynamics, in other words, it does not involve an understanding of what *causes* highly-stacked objects to fall in often cases.

The finding that people may rely on mental shortcuts when facing new situations does not only apply to intuitive physics, similar ideas have also been proposed in the domain of visual concept learning. The prototype theory proposed by Rosch (1975) is a classic example of using heuristics in categorization. Instead of requiring an exhaustive comparison of all attributes of an object with every possible category member, prototype theory posits that people compare new instances to an idealized, average example or "prototype" of a category. This process (also known as the representativeness heuristic) significantly reduces cognitive load by allowing quick, approximate judgments based on the summary statistics of a category rather than on detailed analysis.

0.3.3 The world model hypothesis

At the other end of the spectrum, it is hypothesized that the way people encode past experiences is like building an internal model of the environment which incorporates knowledge about objects, events, and the relationships between them, termed the "world model" (Ullman & Tenenbaum, 2020). This hypothesis is one level higher on the abstraction ladder than the heuristics models (see Figure 1): rather than extracting the summary statistics or relying on a few relevant features from our past experiences, the world model hypothesis argues that we distill the *causal* relationships from the raw inputs and abstract away the unnecessary details, including the features emphasized by the heuristics models, if necessary.

Recently, the world model hypothesis has been supported by various lines of evidence from neuroscience (Rao & Ballard, 1999), psychology (Ullman et al., 2017), and artificial intelligence (Lake et al., 2017) for both visual concept learning and intuitive physics. For example, Lake et al. (2015) showed that participants were able to learn novel hand written characters from just one example and robustly generalize to other instances, and that the one-shot learning behavior can be well captured by a computational model that has access to the underlying causal generative process of these characters. In intuitive physics, the world model hypothesis takes a more concrete form: it is argued that people's ability to understand physical scenes can be explained by having an "intuitive physics engine" in mind. This intuitive physics engine (or IPE), like a game engine, can represent different objects and properties, but runs probabilistic simulations on finite computational resources (K. Smith et al., 2024; Ullman et al., 2017). Evidence in several domains suggests that people's physical predictions can be well explained by the IPE hypothesis, including predicting ballistic motion, judging stability of objects, and even in non-rigid body scenarios (Bates et al., 2015; Battaglia et al., 2013; K. A. Smith et al., 2013; K. A. Smith & Vul, 2013; Wang et al., 2024).

Compared to heuristics models that may lead to systematic deviations from rational judgments, the world model hypothesis can explain how human predictions approximate the ground truth (Sanborn et al., 2013). However, there are also limitations of this hypothesis. One major drawback is the hypothesis's assumption that the mind can construct highly complex and detailed internal models of the world. The computational resources required for such complex models are immense, thus it is questionable how the world model operates efficiently in the mind. In response to this, continued research in this field has proposed resource rational models: a framework that aims to explain human cognition and decision-making as the result of an optimal use of limited cognitive resources. (Griffiths, 2020; Lieder & Griffiths, 2020). For instance, in the domain of intuitive physics, Ullman and colleagues have argued that the mind does not accurately represents the *exact* shape of objects which might be demanding in terms of cognitive resources, but instead utilizes simplification and only represents the approximate contour (Li et al., 2023; Ullman et al., 2017). This is not only much less taxing but also gives good-enough predictions at the same time.

0.4 Current work

Different theories of generalization reviewed above have distinct advantages, and each of them offers explanation to a subset of behaviors *within the same domain*. For example, in the domain of concept learning, the exemplar-based approaches are able explain how people are able to detect and memorize specific units from the inputs in the environment (Fiser & Aslin, 2001; Orbán et al., 2008; Saffran et al., 1996), whereas the prototype theory and the world model hypothesis offer an account of how people are able to extract abstract rule-like patterns (Marcus et al., 1999; Saffran & Wilson, 2003). Furthermore, *across domains*, different theories can explain generalization behaviors to different extent. For instance, the exemplar-based approaches and heuristics models are extensively studied in the domain of visual concept learning, while the heuristics models and the world model hypothesis are mainly explored in the intuitive physics literature.

To what extent should various human generalization behaviors be explained by different domain-specific cognitive mechanisms, or is it possible to be accounted for by a more coherent theory? Both possibilities are plausible: On one hand, it could be that generalization in different domains is indeed supported by different representations and mental processes, just as the literature suggests: memorizing specific units is not the same as learning abstract rules; understanding how two things collide is inherently different from understanding how one object supports another. On the other hand, it is also possible that there are important similarities, allowing for more general-purpose inferential machinery that is applicable across different domains and at the same time can account for different phenomena within each domain as well. The research presented in this dissertation focuses on the question discussed at the beginning: What are the underlying representations and computations that enable people to robustly generalize? I present an approach of designing various scenarios to measure how people generalize to new environments. Leveraging the tools from computational modeling, I explore the possibility that different generalization behaviors across and within domains could be explained by a more unifying account.

Recognizing what entities exist in the world, the concepts, is a prerequisite for richer generalization. As humans, we can readily represent visual structure in our environment — objects, parts, relations, etc. How do people represent regularities in the visual environment? In Chapter 1, I argue that visual concepts are organized as generative models that facilitate generalizing to novel exemplars. I developed an online behavioral experiment where the goal was to learn a novel visual concept. Participants first saw a sequence of images of robots, which were generated according to a set of underlying rules. Participants were then asked to either rate some new robots as how likely they belong to the family of robots they were just shown, or build more robots that belong to the same family. By manipulating the underlying rule, from endorsing learning specific units to learning abstract rules, I find that participants both rated highly and also built novel exemplars that obeyed the same rule as the robots they were shown albeit looking different on the appearance. How are participants able to generalize from the robots they were shown to the novel exemplars they have never seen before? To reverse-engineer the underlying representation of the learned visual concepts, I leverage techniques from program synthesis and build a model that learned from the same data distribution the participants were exposed to and represented the concepts as generative probabilistic programs. I find that across all conditions, participants' responses obtain the highest probability under the model's predictions, compared to alternative models. This provides evidence that both people's memorization of specific units and learning abstract rules can be explained by a program learning model that exploits the statistical regularities in the stimuli, which then enables participants to flexibly adapt to new objects not experienced before.

Being able to generalize over the static entities (e.g. visual concepts) is critical for understanding the environment we live in, but it is not enough to explain how people quickly adapt to novel situations in a *dynamic* world. For example, a novice driver may not have driven over snow before, but they know to drive slowly because they understand that snow can be slippery and send the car into a skid if driving too aggressively. It is not the understanding of individual entities that enables the driver to accomplish this, instead, we need the ability to understand how these entities *interact* with each other in order to make reliable physical predictions about how objects will behave even in novel contexts. How do people learn to "carve physics at its joints" – that is, to uncover hidden variables and rules that can be combined to generalize to new scenarios?

In Chapter 2, I argue that people can combine different physical properties to generalize to novel scenarios and make reliable predictions, and that people's physical predictions is enabled by compositional physical world models. As a starting point, I focus on a specific kind of physical world model, namely models that encode the latent forces and masses of objects in an environment. I develop a novel paradigm where participants were asked to play a physics-based video game. The goal of the game was to catch a ball using a paddle. In order to succeed at this task, participants must infer the existence of different latent forces in different environments, as well as the mass of different balls and compose these to generalize in the test phase. I find that people can learn both latent variables, and critically compose this knowledge to generalize to the novel combinations in the test phase. The modeling results suggest that people achieve such generalization by constructing composable internal models of the physical scene and performing model-based compositional generalization.

Chapter 2 aims to answer a key question in intuitive physics: how do we generalize from our prior experience to make accurate physical predictions in a novel situation that we have never experienced before. However, what happens in everyday life is much richer than that: we do not just generalize to *one* novel scenario, rather, we *constantly* face the problem of generalizing to a variety of different physical contexts. Does the structured model proposed in Chapter 2 support this general-purpose physical reasoning in the real world? Or is it only limited to explaining people's predictions when generalizing to a single physical scenario in a toy physics-based video game? If the latter, what cognitive mechanisms are people relying on when facing different physical contexts? Do people resort to other kinds of representations such as exemplars and heuristics depending on the specific context?

This is the question Chapter 3 aims to address. In Chapter 3, I evaluate human physical predictions on a diverse dataset of rigid body scenarios using naturalistic 3D stimuli. This dataset features a wide range of physical phenomena, each testing some key physical concepts (e.g. collision, support, containment, etc.). Across two experiments, I ask participants to predict whether or where two objects will make contact. I find that participants achieved high accuracy and were highly consistent with each other in their predictions. Moreover, consistent with the world model hypothesis, I show that an intuitive physics engine generalizes to these unseen scenarios in a human-like way, explaining human behavior only slightly worse than expected by the noise ceiling. I also compare models that make heuristic predictions based on initial scene features as well as neural network models. I find that these two kinds of models performed drastically worse than the intuitive physics engine model in terms of explaining human predictions. These results support the world model hypothesis as a generalizable mechanism for intuitive physical reasoning.

In summary, across three chapters, this dissertation shows that people's generalization behavior across and within various domains can potentially be explained by the hypothesis that people have an internal model of the world. In the final chapter, I conclude with a brief discussion of the implication of these findings and directions for future work. Overall, this work lays the foundation for a unifying theory of the underlying representations that enable people to continuously generalize to novel situations.

References

Ambridge, B., Pine, J. M., Rowland, C. F., Chang, F., & Bidgood, A. (2013). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 47–62.

Bates, C., Battaglia, P. W., Yildirim, I., & Tenenbaum, J. B. (2015). Humans predict liquid dynamics using probabilistic simulation. *CogSci*.

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.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition*, *9*(2), 117–123.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, *12*(6), 499–504.

Gilden, D. L., & Proffitt, D. R. (1989). Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 372.

Goldberg, A. E. (2009). The nature of generalization in language.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological review*, *118*(4), 523.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, *24*(11), 873–883.

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.

Harnad, S. (2017). To cognize is to categorize: Cognition is categorization. In *Handbook of categorization in cognitive science* (pp. 21–54). Elsevier.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, *21*(5), 493–518.

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2023). From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*, *130*(4), 977.

Hecht, H., & Bertamini, M. (2000). Understanding projectile acceleration. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 730.

Hecht, H., & Proffitt, D. R. (1995). The price of expertise: Effects of experience on the water-level task. *Psychological Science*, *6*(2), 90–95.

Howard, I. P. (1978). Recognition and knowledge of the water-level principle. *Perception*, *7*(2), 151–160.

Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, *14*, 308–312.

Kozhevnikov, M., & Hegarty, M. (2001). Impetus beliefs as default heuristics: Dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin & Review*, *8*, 439–453.

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & emotion*, *14*(4), 473–493.

Lewis, D. (2008). *Convention: A philosophical study*. John Wiley & Sons.

Li, Y., Wang, Y., Boger, T., Smith, K. A., Gershman, S. J., & Ullman, T. D. (2023). An approximate representation of objects underlies physical reasoning. *Journal of Experimental Psychology: General*.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

McCarthy, W. P., Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *arXiv preprint arXiv:2107.00077*.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, *210*(4474), 1139–1141.

McCloskey, M., & Kohl, D. (1983). Naive physics: The curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 146.

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 636.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, *10*(1), 104.

Nusseck, M., Lagarde, J., Bardy, B., Fleming, R., & Bülthoff, H. H. (2007). Perception and prediction of simple object interactions. *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization*, 27–34.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750.

Paulun, V. C., Bayer, F. S., Tenenbaum, J. B., & Fleming, R. W. (2023). Efficient perception of physical object properties with visual heuristics.

Pavlov, P. I. (2010). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, *17*(3), 136.

Proffitt, D. R., Kaiser, M. K., & Whelan, S. M. (1990). Understanding wheel dynamics. *Cognitive psychology*, *22*(3), 342–373.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87.

Richtsmeier, P., Gerken, L., & Ohala, D. (2011). Contributions of phonetic token variability and word-type frequency to phonological representations. *Journal of Child Language*, *38*(5), 951–978.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*(2), 273–284.

Saltz, E., & Sigel, I. E. (1967). Concept overdiscrimination in children. *Journal of Experimental Psychology*, *73*(1), 1.

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

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic bulletin & review*, *17*(4), 443–464.

Smith, K., Hamrick, J., Sanborn, A. N., Battaglia, P., Gerstenberg, T., Ullman, T., & Tenenbaum, J. (2024). Intuitive physics as probabilistic inference.

Smith, K. A., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. *Proceedings of the annual meeting of the cognitive science society*, *35*(35).

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Son, J. Y., Smith, L. B., & Goldstone, R. L. (2008). Simplicity and generalization: Short-cutting abstraction in children's object categorizations. *Cognition*, *108*(3), 626–638.

Todd, J. T., & Warren Jr, W. H. (1982). Visual perception of relative mass in dynamic events. *Perception*, *11*(3), 325–335.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124–1131.

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.

Wang, H., Jedoui, K., Venkatesh, R., Binder, F. J., Tenenbaum, J., Fan, J. E., Yamins, D., & Smith, K. A. (2024). Probabilistic simulation supports generalizable intuitive physics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Chapter 1

Learning generative models for visual concepts

Abstract

The ability to represent visual structure in the environment — objects, parts, and relations — is a core aspect of human cognition. How do people represent regularities in the visual stimuli? We ran an experiment where participants first saw multiple instances of a novel visual concept, which were generated according to a set of underlying rules. Participants were then asked to either rate as how likely they belong to the family of instances they were just shown (N=163) or generate some new instances that would belong to the same family (N=130). By manipulating the underlying rule, from endorsing learning specific units to learning abstract rules, we found that participants both rated highly and also built novel exemplars that obeyed the same rule as the instances they were shown albeit looking different on the appearance. We further found that across all rule conditions, participants' responses were most consistent with a model that learned from the same data distribution the participants were exposed to and represented the visual concepts as generative probabilistic programs, when compared to alternative models. This provides evidence that both people's memorization of specific units and learning abstract rules can be explained by a program learning model that exploits the statistical regularities in the stimuli, which then enables participants to flexibly adapt to new objects not experienced before.

1.1 Introduction

One fundamental challenge faced by any observer in this world is to interpret their sensory input from the environment and extract meaningful information. Humans accomplish this with great flexibility – we not only can memorize *specific* information encountered in the past, but also represent the regularities in the environment as *abstract* structure that can be then generalized to unseen scenarios (Erickson, 2008; Juslin et al., 2003; Katona, 1940; Murphy, 2004). For example, we can not only remember someone's face, but at the same time we can also effortlessly grasp the correspondence between a real human face and a line drawing of a face (Fig. 1.1), even without auxiliary cues such as color and texture, and categorize them both as face. Moreover, we immediately know that the two dots in the line drawing represents the eyes, the line beneath it represents the mouth, and the big circle represents the head. When do people learn specific instances and when do they represent the regularities as abstract rules?

There has been a long tradition of work investigating how people learn either. Prior research on statistical learning has established that people are able to detect repetition of motifs and represent the statistical regularities in the environment: on one hand, it has been shown that people are able to extract higher-order units out of repeating patterns in both speech (Aslin et al., 1998; Mattys et al., 1999; Saffran, 2001; Saffran, Aslin, et al., 1996; Saffran, Newport, et al., 1996; Saffran & Wilson, 2003; Werker et al., 2012) and visual input (Fiser & Aslin, 2001; Orbán et al., 2008) and represent them as exemplars. For example, it has been shown that people could successfully find artificial word (e.g. *tupiro, golabu, bidaku, padoti*) boundaries in the speech input by relying on the transitional probabilities between syllables. On the other hand, a wealth of research has shown that people are also able to learn algebraic rules (e.g. ABB or ABA patterns like *wo fe fe* or *wo fe wo*) in both modalities (Ferguson et al., 2018; Gomez & Gerken, 1999; Gómez & Gerken, 2000; Marcus et al., 1999; Saffran, 2003; Seidenberg, 1997) when they are endorsed by

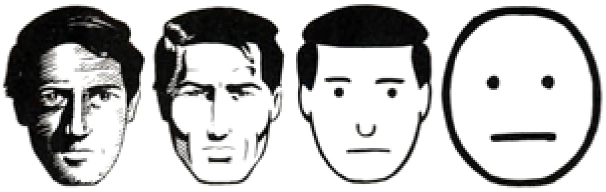

Figure 1.1. Human faces are configured in consistent ways across varying degrees of visual abstraction (McCloud, 1994).

the input statistics. However, an important question left answered is how the statistics of the environment shift people from learning one to learning the other. Answering this question would require careful manipulation of the statistics of the stimuli and investigate the learning behavior under each condition.

A further question revolves around the distinct or shared cognitive mechanisms that support exemplar and rule learning. In the computational modeling literature, there have been divided approaches to explaining how people learn either specific exemplars (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky & Kruschke, 1992) or abstract underlying rules of categories (Bourne, 1974; Koh & Meyer, 1991; Little et al., 2011; Nosofsky et al., 1994; Posner & Keele, 1968). In the rule-learning models, it is often assumed that the participant makes decisions about the values of a stimulus along each of its dimensions of features, and then combines those decisions using logical connectives such as AND, OR, and NOT to determine if the stimulus belongs to a given category. On the contrary, in the exemplar-learning models, researchers usually adopt a similarity-based category representation (e.g. connectionist models), and it is assumed that when asked to generalize to new stimuli, participants pick the training stimulus with the highest similarity to the new stimulus. It remains unclear how these two drastically different kinds of models might be compatible with each other to provide an unifying account for how people are able to learn both exemplars and abstract rules depending on the stimuli. Subsequent work instead adopts a "mixture of expert" model (Anderson & Betz, 2001; Ashby et al., 1998; Bott & Heit, 2004; DeLosh et al., 1997; Erickson & Kruschke, 1998): a hybrid model assuming that both an exemplar-based module and a rule-learning module operate to mediate learning. In these models, one module or the other may predominate for learning the appropriate responses to particular stimuli.

In the current study, we aim to address both questions raised above. In order to investigate how the representations people learn are shaped by the statistics of the environment, we challenged participants to learn novel visual concepts from a sequence of instances and carefully manipulated the statistics of the visual stream. We then investigated the representations participants acquired under different conditions of statistical information by asking them to generalize to novel stimuli. To explain the possible different representations participants may acquired under different conditions, we leveraged techniques from the program synthesis literature in computer science (Gulwani et al., 2017) and explored the hypothesis that people's conceptual knowledge is organized as probabilistic programs – procedures that describe how the observations are generated. One merit of program synthesis make it particular well suited for this purpose: computer programs naturally support higher-order functional abstractions, which can be leveraged to model how perceptual units within objects are be learned dynamically from the visual environment, and how this representation learned adapts to input statistics. Perhaps the study most related to ours comes from Austerweil and Griffiths (2011, 2013), which showed that people infer different feature exemplars to represent the same stimuli depending on the statistical distribution of parts over the objects they observe, and that people's generalization to novel object properties is consistent with predictions from a non-parametric Bayesian model. In this current study, we build on this prior work and explore how people's learning of both exemplars and parameterizable rules can be captured by Bayesian program learning models (Ellis et al., 2023).

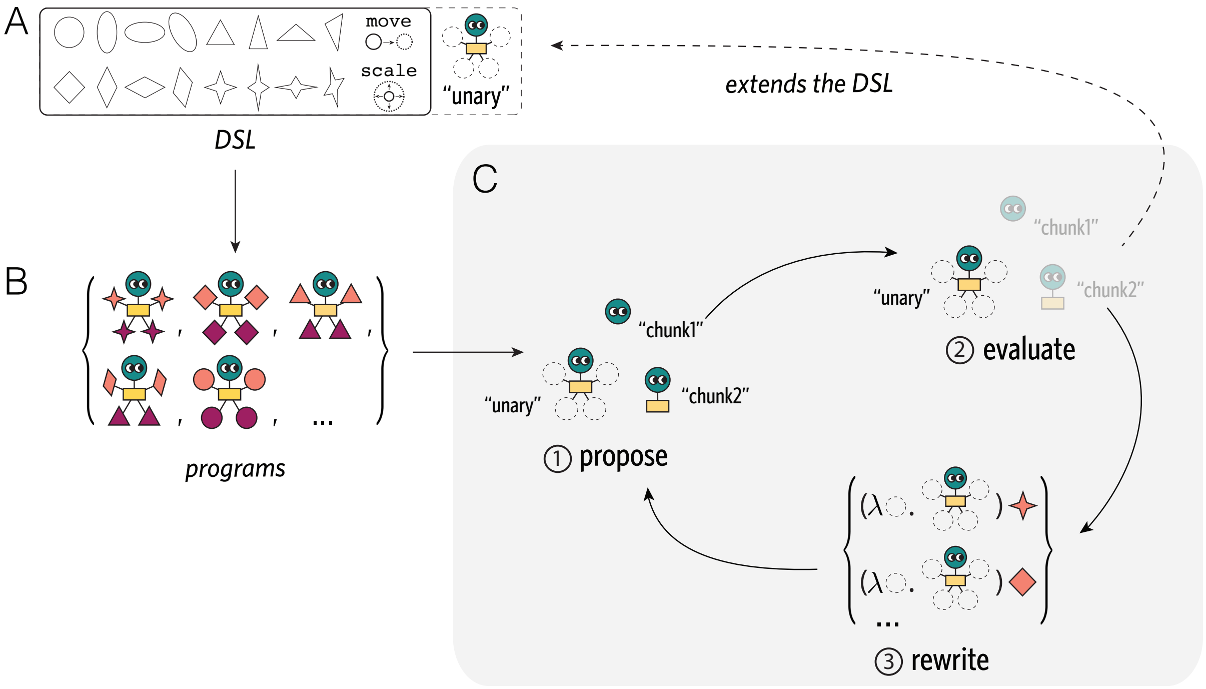

Figure 1.2. Overview of the library learning model. The model takes as input a corpus of robot stimuli, which are represented as graphics programs (B) written in a base DSL (A). Subroutine learning rewrites programs using learned subroutines following an iterative process shown in C. ①: The model first proposes candidate subroutines from the programs, two *chunk* (f^4 and f^7 in text) and a *unary* subroutine are shown as an example. ②: The *unary* subroutine best describes the example robot corpus in terms of posterior probability. ③: The robot programs are re-written using the best-scoring subroutine. See text for more details.

1.2 Computational Model

We formalize our hypothesis in a computational model that leverages programs as representation for visual concepts and uses Bayesian updating for learning a library of program subroutines (Wang et al., 2021). We verify that these mechanisms give rise to the behaviors observed in our empirical data.

1.2.1 Conceptual abstraction as program learning

We begin by specifying how participants' conceptual knowledge is represented and modified over the course of learning in the task. Following Wang et al. (2021), we assume that each participant maintains a library $\mathcal{L}$ of primitives that can be combined to generate simple robots in a domain-specific language (DSL). We assume the library is initialized with 16 primitive shapes, as well as compound shape expressions that represent geometric transformations, and digits `1`∼`9` used as parameters for the geometric transformations (see Fig. 1.2A, the digits are omitted for brevity). This DSL is small but fully expressive: any possible robot can be represented as a program by combining together these basic commands. More generally, each robot consists of six features: head, torso, left and right hand, and left and right foot; out of these features, the head is always a `circle`, the torso is always a `rectangle`, and the other features can manifest as any primitive shape. Fig. 1.2B depicts five different robots expressed in the DSL, the corresponding programs for the first two robots are shown below:

```
p1 = set(move(2.0, 1.0 (scale(2.0, star))),
         move(4.0, 1.0 (scale(2.0, star))),
         move(3.0, 2.0 (scale(2.0, rectangle))),
         move(1.0, 3.0 (scale(2.0, star))),
         move(5.0, 3.0 (scale(2.0, star))),
         move(3.0, 4.0 (scale(3.0, circle))))
```

$p_2 =$

```
set(move(3.0, 2.0 (scale(2.0, rectangle))),
    move(1.0, 3.0 (scale(2.0, square))),
    move(5.0, 3.0 (scale(2.0, square))),
    move(2.0, 1.0 (scale(2.0, square))),
    move(4.0, 1.0 (scale(2.0, square))),
    move(3.0, 4.0 (scale(3.0, circle))))
```

As can be seen, these programs are "flat" and lack any structure: each program is simply a set of primitive shapes, and each shape is individually moved and scaled to appear at the right position.

In the Bayesian program learning framework, the DSL is updated over time by expanding the library with new primitives. As a participant progresses through multiple trials of robots $\{R_n\}_1^N$, they may extract common subroutines that would allow them to *re-represent* the data more efficiently (see Fig. 1.2C). Consider the two programs shown above as an example: they share the same structure in the sense that they place six body features at the same positions, but they differ in the shapes of hands and feet. The model represents this common structure and variation explicitly by rewriting the corpus of programs using a library of learned subroutines. Here the learned subroutine `f(x)` has the same structure as p_1 and p_2, but the concrete shapes of hands and feet are replaced with a parameter, `x`;

```
f(x) = set(move(3.0, 2.0 (scale(2.0, rectangle))),
           move(1.0, 3.0 (scale(2.0, x))),
           move(5.0, 3.0 (scale(2.0, x))),
           move(2.0, 1.0 (scale(2.0, x))),
           move(4.0, 1.0 (scale(2.0, x))),
           move(3.0, 4.0 (scale(3.0, circle))))
```

with this subroutine in hand, we can rewrite the two programs more compactly by simply making a call to this subroutine with different arguments: `f(star)` and `f(square)`, respectively (also see ③ in Fig. 1.2C for visual illustration):

$$
\begin{aligned}
p'_1 &= \texttt{f(star)} \\
p'_2 &= \texttt{f(square)}
\end{aligned}
$$

Note that this representation is more *compact*: while the total size of p_1 and p_2 is $37 \times 2 = 74$ tokens, after the rewriting, the total size of `f`, p'_1 and p'_2 is only $38 + 2 + 2 = 42$ tokens (the size of `f` is the size of its body plus the number of parameters). This compression of the dataset is achieved thanks to a careful choice of the subroutine, which maximally captures the common structure between p_1 and p_2 and abstracts away the differences.

The model learns a library of subroutines by repeatedly performing the following three steps (Fig. 1.2C): (1) proposing *fragments* from the programs as candidate subroutines; (2) evaluating these candidates according to the Bayesian posterior; and (3) re-writing all programs using the highest scoring candidate and adding it to the library.

In the rest of this section, we describe the three steps in more detail.

Propose

In the first step, the model proposes a set of candidate subroutines by extracting fragments from the given programs, matching those fragments against fragments of other programs in the dataset, and abstracting away their differences into parameters using a technique known as *anti-unification* (Plotkin, 1970). Let us illustrate this procedure on our running example. Given the program p_1, we first generate a fragment for each of the 64 subsets of its shapes; for example, the following fragment f^4 represents just the head and f^7 represents head and torso (see ① in Fig. 1.2C):

$$f^4 = \texttt{move(3.0, 4.0 (scale(3.0, circle)))}$$

$$f^7 = \texttt{set(move(3.0, 2.0 (scale(2.0, rectangle))),}$$

$$\texttt{move(3.0, 4.0 (scale(3.0, circle))))}$$

Finally $f^{64} = p_1$ is a fragment representing the whole robot.

Next, the model attempts to match (or *anti-unify*) each fragment f_i^k from p_1 with each fragment f_j^m from p_2; for the match to be successful, we require that the numeric parameters of `move` and `scale` must coincide, while the primitive shapes in the two fragments might differ, in which case they are replaced by a parameter. For example, matching f^4 with the *torso* fragment from p_2 fails, since their positions differ. Matching f^4 with the *head* fragment from p_2 succeeds and yields f^4 itself as a candidate subroutine, since the two fragments are identical; this subroutine is a *chunk*, i.e. has no parameters. Finally, matching f^{32} against the whole p_2 succeeds and yields a *unary* subroutine `f(x)`, i.e. a subroutine with a single parameter, x, which replaces the mismatched shapes (see ① in Fig. 1.2C for a visual illustration of *chunk* and *unary* subroutines). This subroutine is unary because all mismatches between the two fragments are of the same form: `star` for p_1 vs `square` for p_2. More generally, matching two incongruent robots (where the shape of the hands differs from the shape of the feet), may yield a *binary* subroutine `f(x,y)` with two parameters: one for the shape of the hands and the other one for the shape of the feet. In total, a dataset with just $\{p_1, p_2\}$ yields 64 candidate subroutines, because each fragment of p_1 has a unique matching fragment in p_2. For a larger dataset, the number of candidates can grow.

Evaluate

Formally, the model updates a posterior distribution over possible ways of extending the library using the proposed subroutines (including $f = \emptyset$, which would maintain the current library):

$$P(\mathcal{L} \cup \{f\} | \{R_n\}_1^N) \propto \underbrace{P(\mathcal{L} \cup \{f\})}_{\text{description-length prior}} \times \underbrace{\prod_{n=1}^{N} P(R_n | \mathcal{L} \cup \{f\})}_{\text{likelihood}} \tag{1.1}$$

This posterior distribution weighs two competing criteria for a good library: the likelihood and the prior. The *likelihood* in Eq. 1.1 captures the ability of an extended library efficiently to explain previous towers:

$$P(R_n | \mathcal{L} \cup \{f\}) = \mathsf{exp}(-\mathsf{MDL}(R_n \mid \mathcal{L} \cup \{f\})) \tag{1.2}$$

where MDL is a function evaluating the *minimum description length* of a program given a library, defined as:

$$\mathsf{MDL}(p \mid \mathcal{L}) = \min\{\mathsf{size}(p') \mid p' \overset{\mathcal{L}}{\to}{}^{*} p\}$$

Here $p' \overset{\mathcal{L}}{\to}{}^{*} p$ means that the program p' can be rewritten into p by substituting calls to subroutines from $\mathcal{L}$ with their definitions. Intuitively, the MDL is the most compact version of R_n that can possibly be written in the updated library $\mathcal{L} \cup \{f\}$. This term is therefore maximized by sets of fragments $\{f\}$ that allow the existing data to be expressed most efficiently.

The *prior*, on the other hand, captures a preference for smaller libraries:

$$P(\mathcal{L} \cup \{f\}) = \mathsf{exp}(-w \cdot \mathsf{size}(\mathcal{L} \cup \{f\})) \tag{1.3}$$

where $\mathsf{size}(\mathcal{L} \cup \{f\})$ represents the number of primitives in the updated library. The strength of this preference is controlled by a parameter w, reflecting the cost of learning. We explore several values of w in our simulations. Intuitively, when $w = 0$, there is no penalty for having a larger library, so the model tends to learn larger libraries. As $w \to \infty$, any expansion of the library is considered too costly, preventing library learning entirely. These two objectives balance out in the posterior distribution (Eq. 1.1) such that the fragments f with the highest posterior probability are those that provide maximal compression of input robot programs while minimizing expansion of the library.

In practice, we selected the single highest posterior-probability set of fragments in each iteration of the loop in Fig. 1.2C. In our running example, among the 64 candidates proposed for the dataset $\{p_1, p_2\}$, the highest-scoring candidate is the unary subroutine `f(x)`, which represents the whole robot (see ② in Fig. 1.2C).

Rewrite

After the candidate subroutine with the highest posterior probability has been selected and added to the library, all current programs are re-written to make use of the new subroutine. For example, as shown in Fig. 1.2C ③, the programs p_1 and p_2 are rewritten in terms of the newly added subroutine `f(x)`. The resulting programs are used as the input to the next iteration of library learning. The process stops when none of the proposed candidate subroutines improve the loss any further. In our example, p_1' and

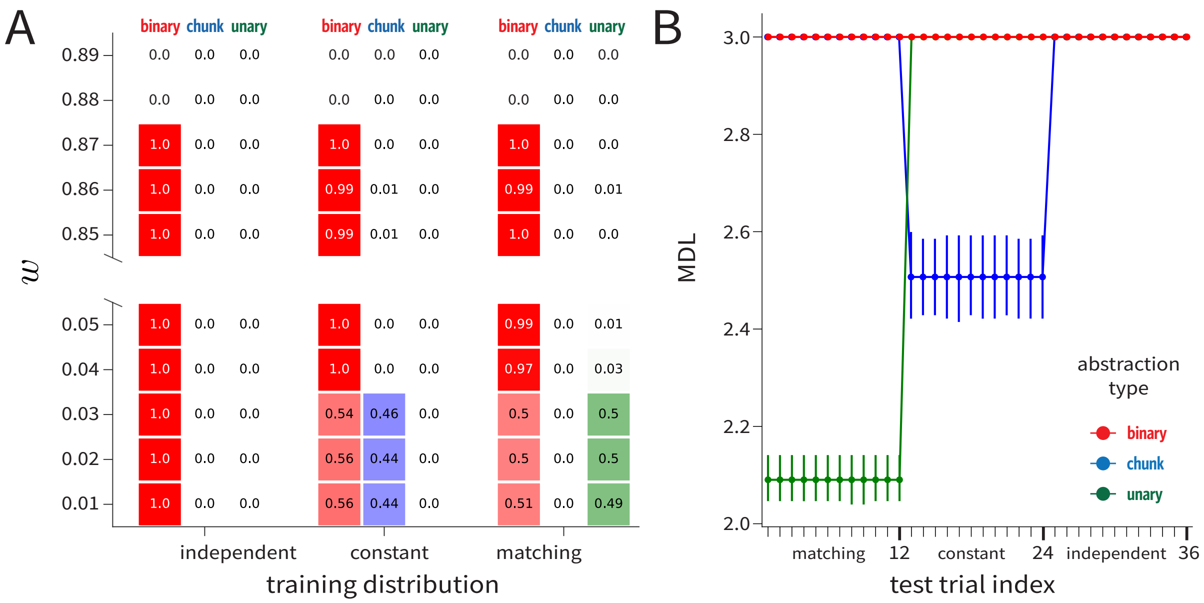

Figure 1.3. (A) Heat map representing the relative frequency with which different kinds of subroutines (i.e. chunk, unary, and binary) were learned, for different values for the cost of learning (y axis) and different covariance between features (x axis). (B) The minimal description lengths of each robot given by models trained on different distributions ($w = 0.02$). Different colors represent models learned on different distributions. The independent, matching, and constant test trials are grouped together for easier visualization (x axis). Error bars represent 95% CIs. Notice how conditioned on the training data distribution, the models would evaluate the same robots during the test phase differently (for a particular robot, the three colored lines are different); and that the same model would also evaluate rule-consistent and rule-inconsistent robots differently during the test phase (for a particular colored line, the MDLs for different robots are different).

p_2' each contain only a single fragment, and the result of matching these fragments is a unary candidate subroutine `g(x) = f(x)`. Rewriting p_1' and p_2' in terms of `g(x)` does not change their size (it simply replaces `f` with `g`); hence this candidate is rejected, since it increases the size of the library without compressing the corpus.

1.2.2 Simulation Result

The goal of our simulations was to explore how different constraints on learning, supplied either by external variables (i.e. the robot distribution the model was trained on) or internal components of the model (i.e. w in Eq. 1.3), jointly influenced the subroutines the model learned. We also show how the effect of learning different subroutines was manifested when we asked the models to evaluate different robots during test. The evaluations made by the models in this section will be directly tested against participants' predictions in the next section.

Experiment setup

Building on classic work investigating perceptual learning in cognitive science (Austerweil & Griffiths, 2013; Goldstone, 2003), we designed our simulation experiments to vary along two dimensions:

- **Cost of learning**: We varied the value of w, the parameter controlling the preference for smaller libraries during learning, from 0 to 1. We hypothesized in the previous section that as w becomes higher, the model is pressured to learn a more compact abstraction of the training distribution and hence could be reused across a larger proportion of the data. When w is very high, any expansion of the library is considered too costly, preventing library learning entirely.
- **Covariance between features**: We varied how strongly particular robot body features co-occurred. Specifically, we designed three different distributions of robots used for training. In the *matching* distribution, hands and feet of the robots always take the same shape. In the *independent* distribution, shapes of hands and feet are independent. In the *constant* distribution, shape of hands (or feet) never vary, and shapes of feet (or hands) is sampled independently from the shape of hands (or feet). In all three distributions, the shapes of the two hands or two feet are the same.

We generated 100 samples from each of the three data distributions using 12 of the 16 shapes from the DSL. Each sample contained 200 robots, with 10% noise (e.g. for the constant distribution, 180 out of the 200 robots had the same shape for hands, the other 20 robots' hands were independently sampled). We then ran the model for each sample at different w and analyzed the library of learned subroutines. To measure how the data distribution and cost of learning affect the level of abstraction of the learned concepts, we analyzed whether each library contained a "robot subroutine" (i.e. a subroutine with a head, torso, hands and feet) with different number of parameters:

- *Chunk* subroutine: these subroutines contain constant shape primitives for hands or feet, reflecting cases where the model does not learn more abstract structure but instead memorizes subsets of the original training programs. In the matching and independent distribution, these subroutines tend to only be applicable to a small number of robots, and thus not reusable across examples. In the constant distribution, however, since most of the robots in the training distribution have the same shape primitive for some of their body features, we expect the model to extract these commonalities by learning the chunk subroutine.

- *Unary* subroutine: these subroutines have a single parameter, and hence are able to abstract over a single shape inside the robot (e.g. hands and feet in a robot). This kind of subroutine is more expressive than the constant subroutine because it can be applied to new shapes that the model has never seen before so long as they exhibit the same structure. However, it is also limited in the sense that it can only describe robots whose hands and feet share the same shape, thus it is not a reasonable abstraction of the independent distribution.

- *Binary* subroutine: this subroutine has two parameters, one for the the hands and one for the feet. It is the most flexible abstraction, since it captures all variations of body features in the training distribution.

Subroutines learned under different conditions

First we explored the subroutines learned by the model when given different training distributions. We found that, consistent with our hypothesis, when provided with the constant and matching distribution, the model indeed learned the chunk and unary subroutine, respectively (see Fig. 1.3A). Interestingly, we also noticed that the binary subroutine was usually learned along with the chunk or unary subroutine. This is because, as mentioned before, the binary subroutine is the most flexible abstraction thus can be used as the default representation of the robots in the face of noise in the training distribution. In the independent data distribution, it is no surprise that the model only learned the binary subroutine as there is no correlation between the shapes of the hands and feet.

We next explored the consequences of varying w, the cost of learning, over a range, $0 \leq w \leq 1$. As can be seen from Fig. 1.3A, at a low level of w ($w = 0.01$), the model was free to learn different types of subroutines that can both exploit the regularities in the training distribution, and at the same time flexibly accommodate noise. As w increases, it became harder to learn distribution-specific subroutines, thus forcing the model to fall back to the most flexible, distribution-agnostic abstraction – the binary subroutine. And at a high level of w ($w = 0.88$), it becomes too expensive to learn any subroutine, preventing library learning entirely.

Predictions on generalization

So far we have seen that under different external and internal constraints, our model is able to exploit statistical regularities in the data distribution and learn abstractions that are reusable. However, in order to test the validity of our model, it needs to make predictions on behaviors that we can directly test against participants' data. So in this section we ask: what do the learned subroutines entail? One natural way to answer this question is to look at how the model evaluates different robots using the subroutines it learned.

In order to do this, we looked at a w level where various subroutines can be learned ($w = 0.02$) and randomly sampled a model in each of the three training distributions to form a triplet, resulting in 100 triplets. In so far as the model can learn abstractions tailored to its training distribution, then when presented with robots vastly different from its training distribution, the model will find them hard to describe using the subroutines it has learned. This implies when presented with the same set of robots, the models from the triplet should evaluate them differently. For example, when presented with a set of robots whose shapes of hands and feet have no correlation, the model trained on the independent distribution will find them natural using the binary subroutine it learned. However, these same robots will be deemed highly unlikely by the models trained on the matching and constant distribution, because they cannot be easily explained using the learned unary and chunk subroutine. Following this idea, in practice, for each triplet we sampled 36 robots as *test* trials, these 36 trials not only included robots that used the same 12 shapes for hands and feet as in training, but also new robots using the other four shapes. To increase the variability in the sets of robots evaluated by the triplet of models, we designed the test trials such that they consisted of 12 robots compatible with each of the the unary, chunk and binary subroutine (see the x axis of Fig. 1.3B). We presented the same 36 robots to each model in the triplet and calculated the MDL (minimum description length, see Eq. 1.2) of each robot using the model. We do so for all 100 triplets and the results are shown in Fig. 1.3B.

As we can see, the models learned on the three training distributions indeed made very different judgements. The models that learned the unary subroutine on the matching distribution (green line in the figure) can represent the matching robots (test trial index $1 \sim 12$) in the test trials compactly (low MDL), but not other robots. The models that learned the chunk subroutine on the constant distribution (blue line in the figure) can represent the constant robots (test trial index $13 \sim 24$) in the test trials compactly (low MDL). The models trained on the independent distribution, represented by the red line, learned the binary subroutine. And since the binary subroutine is the most flexible and can describe any robot, it is not surprising that it gave the same MDL for all test robots.

1.3 Experiment 1

An online experiment was designed to compare our model of learning visual concepts with people's judgments on a range of "robot learning" scenarios, the task introduced in previous sections.

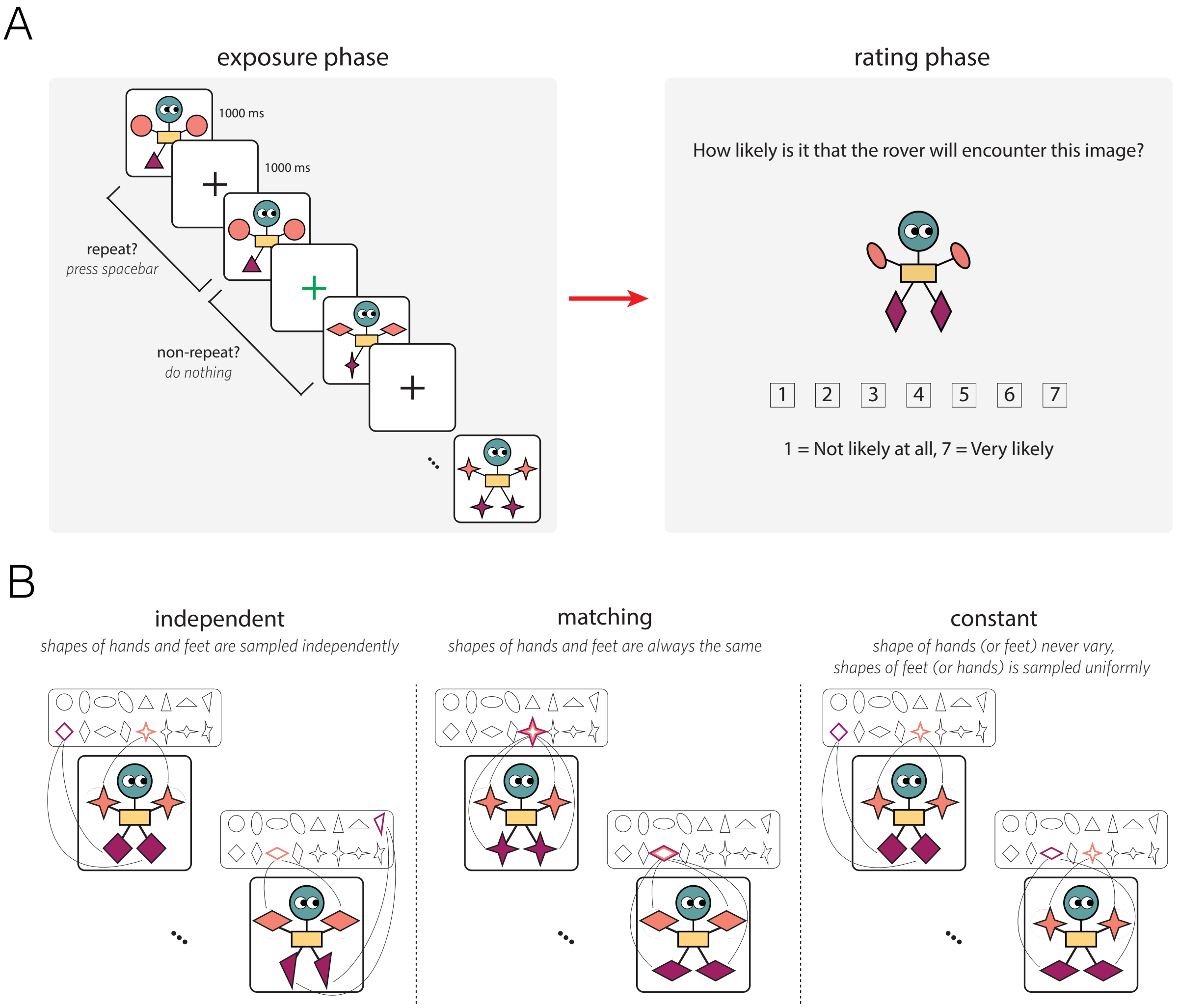
A
exposure phase
1000 ms
1000 ms
repeat?
press spacebar
non-repeat?
do nothing
rating phase
How likely is it that the rover will encounter this image?
1 2 3 4 5 6 7
1 = Not likely at all, 7 = Very likely
B
independent
shapes of hands and feet are sampled independently
matching
shapes of hands and feet are always the same
constant
shape of hands (or feet) never vary,
shapes of feet (or hands) is sampled uniformly

Figure 1.4. (A) The procedure of experiment 1. During training, participants were exposed to a sequence of robots and were asked to press the space bar when they saw a repeat. They were then asked to rate how likely they would see a set of robots during the test phase. (B) The design of experiment 1. Participants were randomly assigned to three training distributions: independent, matching and constant where the rule for the shapes of robot hands and feet differed.

1.3.1 Participants

188 participants were recruited from Prolific. Data from 25 participants were excluded for failing our preregistered inclusion attention checks. Participants provided informed consent in accordance with the institution's IRB. The experiment lasted approximately 15 minutes and participants were paid $14/hr based on this expected completion time.

1.3.2 Stimuli

To understand how people learn and represent novel concepts, we used the same robots used in the last section for the models as stimuli for the human experiment. As introduced before, the shapes for head and torso of the robots never changed, whereas the shapes for hands and feet were sampled from a distribution. 16 shapes were designed for the hands and feet and they varied on two different axes: the shape identity (circle, square, triangle and star) and the ratio of the shape (isomorphic, wide, tall and slanted) resulting in 16 (4×4) different shapes (see Fig. 1.4B for the 16 shapes), these were the same shapes we used in the DSL for the models.

1.3.3 Design

To probe what was learned by the participants in the experiment, mirroring what we did for the models, we trained the participants on a subset of the 16 shapes and then asked them to generalize to the rest during test. We split these 16 shapes into training and test by the ratio axis, where three out of the four ratios were used for training and the held-out ratio for test, resulting in 12 out of 16 shapes used for hands and feet during training as in the model simulations. The assignment of split along the ratio axis were randomized across participants. Like in the model simulations, we designed three different data distributions to elicit participants' different hypotheses about the robots: matching, independent and constant (see Fig. 1.4B for visual illustration). Participants were randomly assigned to one of these three conditions. For the constant condition, either the hands or the feet could take the constant shape and the other varies, to balance this, participants were randomly assigned to one of the two hands or feet groups, and the choice of the constant shape was randomly sampled from the 12 training shapes.

To keep participants engaged, the training phase was structured as a one-back task, where the participants were told to monitor a stream of robots and to press the space bar whenever they observe an immediately subsequent repetition of a robot in that stream. Each participant received 200 trials during training with 20% noise, i.e. 160 rule-consistent trials and 40 rule-inconsistent (noise) trials.

In section 1.2.2, we predicted that conditioned on the data distribution participants are trained on, they will evaluate the same robots differently during the test phase; and that the same participant will also evaluate rule-consistent and rule-inconsistent robots differently during the test phase. To test this hypothesis, we asked the participants to give likert scale ratings to a set of robots during test to indicate how likely they think they would see these robots. More specifically, we designed a yoked test phase, where a triplet of participants trained on the matching, independent and constant condition were asked to rate the same robots during the test phase, mirroring what we did in section 1.2.2. The test trials consisted of three groups: 12 robots that were consistent with the matching condition, eight from the 12 matching robots were sampled using the 12 shapes participants saw during training and four robots using the four new shapes saved for test; 12 robots that were consistent with the constant condition, using the same sampling procedure as the independent test trials; and 12 robots that were consistent with the independent condition. This resulted in 36 test trials in total.

1.3.4 Procedure

Before the training phase, the participants were given the instructions below:

Recently a Mars rover found relics of humanoid robots in a cave. A team of *scientists believes these robots might be left over from an advanced civilization. Scientists are studying these robots more closely to learn more about them. Your job is to help these scientists learn the visual characteristics of these robots. Specifically, they need your help understanding what features they share and which features vary between individuals. Here is how it is going to work: First, you will view a set of examples that the scientists will present to you, one by one. Please press the space bar when you see a robot that is the same as the previous one. Then you will be asked to answer some questions about some other robots based on the patterns you notice.*

They were then presented with robots shown sequentially (1000ms), with a fixation cross interleaved (1000 ms). The participants were asked to press the space bar when they see a repeat, and then the fixation cross would turn green regardless whether their response is correct or not (see Fig. 1.4A, left). 20 repeated trials were inserted into the sequence intentionally every 10 trials. The exact location of the repeat in these 10 trials was jittered. The full training sequence was divided into two blocks, 100 trials each with a short (2 min) break in the middle.

On the test trials, participants were presented with 36 robots and were asked to rate how likely they believe it is that the rover sees each image as it explores further through the cave, on a $1 \sim 7$ likert scale (see Fig. 1.4A, right). The instruction is shown below:

It looks like there are many more images on the cave wall that the rover has not yet had a chance to record. If the rover explored the cave wall further, which images do you think it would be likely to see? Your task is to rate how likely you believe it is that the rover sees each image as it explores further through the cave.

1.3.5 Results

We now turn to a comparison of participants' behavior with the Bayesian program learning model introduced previously. In section 1.2.2, we showed that our model could

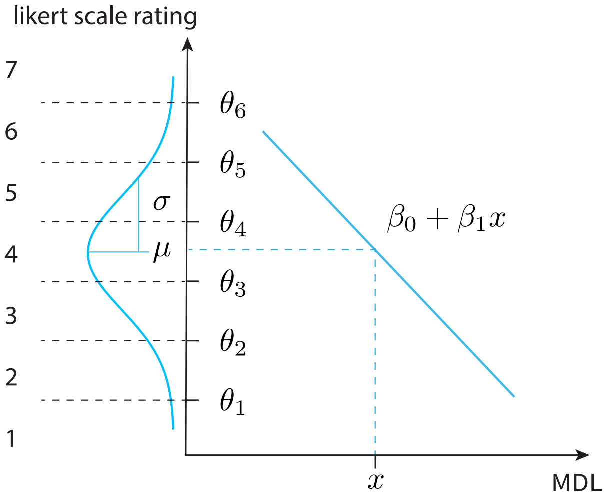

Figure 1.5. Using the Probit model as the likelihood function: the x axis is the MDL calculated by the library, the y axis is the $1 \sim 7$ likert scale ratings. The linear function $\beta_0 + \beta_1 x$ specifies the mapping between the two axes. The free parameters of the Probit model are β_0, β_1, σ and $\theta_{1\sim 6}$.

make quantitative predictions of what participants would do when presented with a set of robots during the test phase. In this section, we will show that the likert scale ratings we collected from the participants can be well explained by our model.

Specifically, we calculated the likelihood of a library under the collected participants' ratings as follows:

$$P(\{r_n\}_1^{36}|\mathcal{L}) = \prod_{n=1}^{36} P(r_n|\mathcal{L}) \tag{1.4}$$

where r_n is the participant's rating $(1 \sim 7)$ on the nth test trial, $\mathcal{L}$ is a learned library. On any given test trial, a MDL was calculated by the library and a rating was given by the participant. To map the library's predictions (MDL) to likert scale ratings on each test trial, we specified $P(r_n|\mathcal{L})$ as a Probit model (Bliss, 1934; Gelman et al., 1995). The Probit model assumes there is some noisy mapping from the library's prediction (MDL) to the participant's response scale, and the probability of each discrete likert scale response is the area under the normal curve between the thresholds for that response

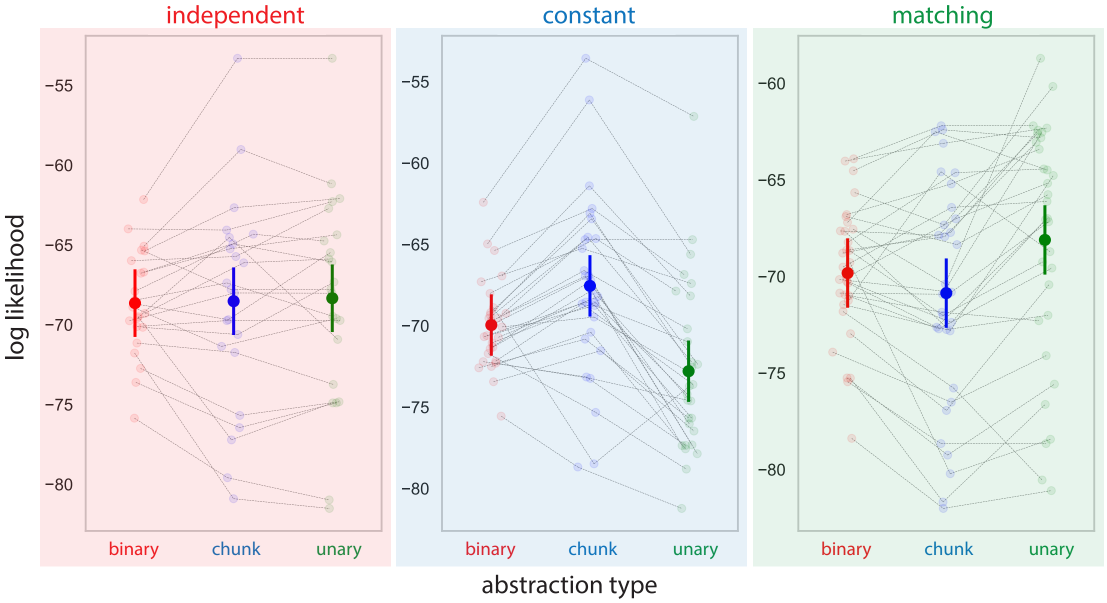

Figure 1.6. Log likelihoods of different library learning models under different training distributions in experiment 1. Each colored panel represents a training distribution (e.g. "independent", "constant" and "matching"), each connected line represents the log probabilities of *one* participant's ratings scored under different library learning models (e.g. "binary", "chunk" and "unary"). Error bars represent 95% CIs.

category (see Fig. 1.5). The free parameters for the Probit model are the slope (β_1) and intercept (β_0) of the linear mapping function and the parameters of the Probit observation function (the standard deviation σ, and the thresholds θ_i).

We hypothesized that, in so far as participants can exploit the statistical regularities in the training distribution, that conditioned on the data distribution participants were trained on (independent, matching and constant), a triplet of participants would give heterogeneous ratings to the same robots during the test phase (see Fig. 1.3B). Therefore, participant's ratings should be endorsed most by the model trained on the same data distribution, compared to models trained on other distributions. To test this hypothesis, given the ratings from a participant trained on one distribution, we not only calculated the likelihood of the library learned from the same distribution ("congruent library"), but also calculated the likelihoods of the libraries learned using the other two alternative training distributions (“incongruent library”) in the triplet for comparison. We predict that the congruent library would get the highest likelihood compared to the libraries trained on the other two data distributions.

We fitted the parameters of the Probit model $(\sigma, \theta_i, \beta_0, \beta_1)$ together with the free parameter of the Bayesian program learning model w (“cost of learning”, see Eq. 1.3) using half of participants’ data and evaluated the likelihood of the different libraries using the other half, shown in Fig. 1.6. It can be seen that the log-likelihoods of the congruent libraries are significantly higher than the incongruent libraries. Specifically, we first fitted a linear mixed-effects model to predict the probability of participants’ likert scale ratings during test from training distribution type and library type, with random intercepts for each participant. We then added an interaction term between training distribution type and library type and found that this model significantly outperformed the base model without interaction term ($\chi^2 = 67.0$, $p < 0.001$)[1], suggesting that the effect of the library type differed for different types of training distribution (see Fig. 1.6). But what form does that difference take? We further fitted a linear contrast model where we compared the log-likelihoods of congruent libraries versus incongruent libraries across different types of training distributions. We found strong evidence that the congruent libraries outperformed the incongruent libraries ($p < 0.01$).

1.4 Experiment 2

In the previous section, we showed that participants’ likert scale ratings can be well explained by the learned libraries. Although this provides evidence for the possibility that people’s conceptual knowledge is organized as programs, it is not clear that such structured representation is *required* for such simple, implicit discrimination tasks (i.e. likert scale ratings). Can we design stronger tests that directly probe the underlying representation of

[1] This effect was reliable across 100 random splits of participants’ data for estimating parameters and evaluating likelihoods.

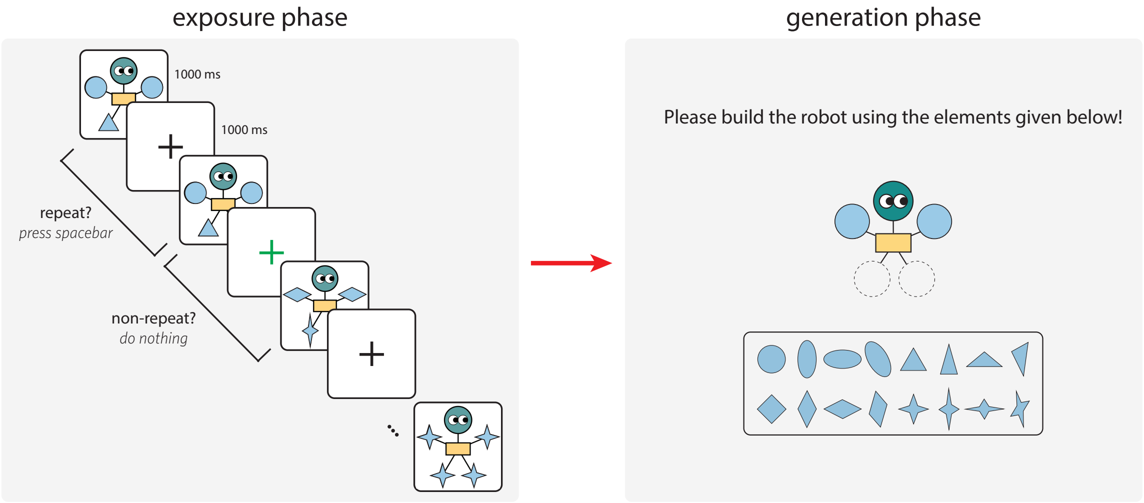

Figure 1.7. The procedure of experiment 2. The training phase was the same as experiment 1 except for the color of robot hands and feet being the same. In the test phase participants were asked to build robots with either hands or feet filled in for them using shapes from provided below.

visual concepts in a more explicit way? In this section, we conducted a second experiment where we asked the participants to *build* robots in the test phase.

1.4.1 Participants

144 participants were recruited from Prolific. Data from 14 participants were excluded for failing our preregistered inclusion attention checks. Participants provided informed consent in accordance with the institution's IRB. The experiment lasted approximately 15 minutes and participants were paid $14/hr based on this expected completion time.

1.4.2 Stimuli

We used the same stimuli as in experiment 1, the only difference is that we used the same color for hands and feet of the robots (see Fig. 1.7). This is due to the consideration of the experiment interface during the test phase: although useful for segmenting hands from feet during training, the different colors for hands and feet posed a challenge for the response options to be consistent across trials in the test phase (e.g. when asked to fill in the hands, the participants see orange shapes options; while asked to fill in the feet, the participants see purple shapes options). So in experiment 2, we chose the same color for hands and feet of the robots, even at the cost of making the learning in the training phase harder.

1.4.3 Design

The training phase was the same as experiment 1. For the test phase, we predicted that conditioned on the data distribution participants were trained on, participants would build different robots when given the same constraints. To test this hypothesis, similar to experiment 1, we designed a yoked test phase, where a triplet of participants trained on the matching, independent and constant condition were asked to build some robots given the same constraints.

Specifically, test-phase consist of two types of trials: three trials on which robots already had hands and three trials where the robots already had feet. The purpose of having these two types of trials was because they help to distinguish the predictions made by the three concept libraries. For example, in the yoked triplet, if the participant trained on the "constant" distribution was trained on a distribution where all robots had circular hands, then the six test trials would consist of: (a) three trials with circular hands filled in for them and (b) three trials with feet filled in for them using shapes randomly sampled from the other 15 shapes, ensuring that one of these three trials was using a novel shape (and thus enabling participants to generalize to novel examples following the same pattern). These six trials would be the same for the participant trained on the "matching" distribution and the participant trained on the"independent" distribution in the yoked triplet. On each trial, the participants were given all 16 shapes as a library they can choose from (see Fig. 1.7). The sequence in which these six test trials are presented were randomized for each participant.

We hypothesize that for the six trials, participants in the "mathcing" group would always select the congruent shape, regardless of whether they have seen that particular shape before or whether it's hands or feet being filled in for them. Participants in the "independent" group would always randomly select one of the 16 shapes given to them, regardless of whether it's hands or feet being filled in for them. For participants in the "constant" group, for trials where the constant body part they were trained on has already been filled in for them (circular hands in the example above), participants would randomly select one of the 16 shapes given to them to fill in the body part they were not trained on (feet in the example above), however for trials where the constant body part they were trained on is empty, participants would select the shape they were trained on to fill in the empty slots.

Furthermore, since the task of experiment 2 was more demanding and required more explicit decisions and choices during the test phase, aside from conducting the experiment at the same noise level as experiment 1 (20%), we also conducted the experiment at a lower noise level (7%) where the regularities were more substantial.

1.4.4 Procedure

The procedure is the same as experiment 1, except for the task instructions given before the test phase:

It looks like there are some robots in the cave that are broken. Your job is to repair these robots to bring them into working condition. Notice that in order for any of these robots to work, as a group they will need to look representative of the robots that you have seen so far. So please do your best to repair these broken robots such that they will reflect the similarities and differences that you noticed among the robots you have seen so far.

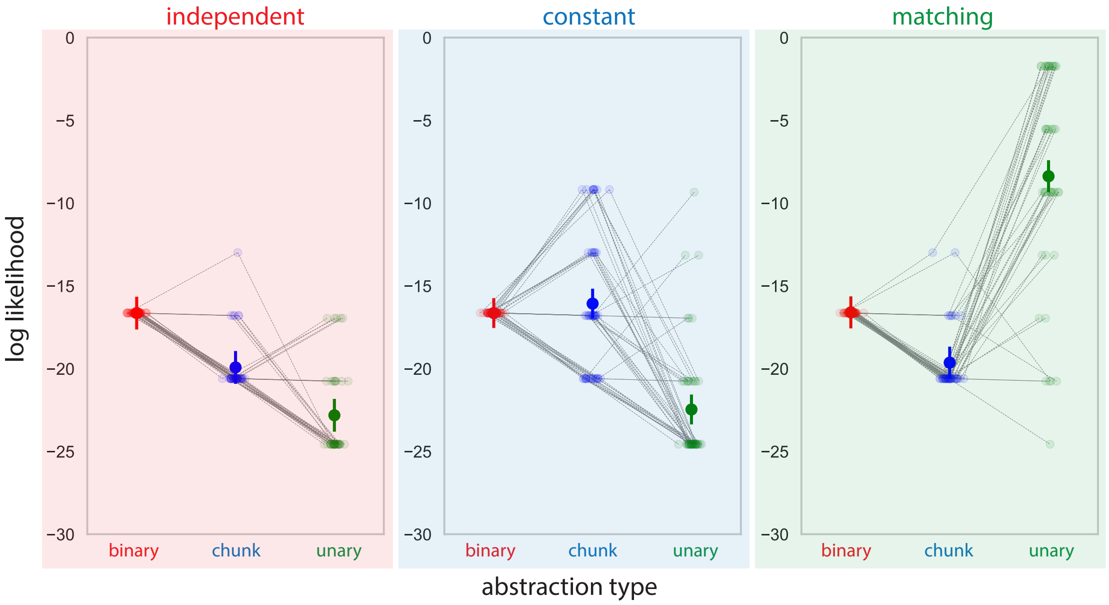

Figure 1.8. Log likelihoods of different library learning models under different training distributions in experiment 2. Similar to Figure 1.6, each connected line represents the log probabilities of the robots built by *one* participant scored under different library learning models (e.g. “binary”, “chunk” and “unary”).

1.4.5 Results

Similar to experiment 1, we compared participants’ behavior with the Bayesian program learning model by calculating the likelihoods of different libraries that can be learned:

$$P(\{b_n\}_1^6|\mathcal{L}) = \prod_{n=1}^{6} P(b_n|\mathcal{L}) \tag{1.5}$$

where b_n is the robot built by the participant on the nth test trial. We specified $P(b_n|\mathcal{L})$ as a binomial distribution where the model had probability p of building robots that are consistent with the learned library and probability $1-p$ of building other robots, reflecting noise.

Similar to experiment 1, we fitted the parameter p and w using half of participants’ data and evaluated the likelihoods of different libraries using the other half, shown in

Fig. 1.8. As can be seen, the robots participants built during the test phase obtained the highest probability under the type of library learned from the distribution that is the same as the data distribution participants were trained on. Specifically, we fitted a linear mixed-effects model to predict the probability of participants' generated robots during test from training distribution type, library type and noise level, with random intercepts for each participant. We then added an interaction term between training distribution type and library type and found that this model significantly outperformed the base model without interaction term ($\chi^2 = 302.5$, $p < 0.001$), suggesting that the effect of the library type differed for different types of training distribution. By fitting a linear contrast model, we found that the difference was because congruent libraries outperformed the incongruent libraries ($p < 0.001$) across noise levels. And moreover, this was the case for both groups of participants at different noise levels ($p < 0.001$ for both 20% and 7% noise level group).

1.5 Discussion

In this paper, we investigated how the representations people learned are shaped by the statistics of the visual environment. Specifically, we developed a study where participants were first exposed to a sequence of robot images. We manipulated how the shapes of different body features co-occurred such that they followed different statistical patterns. In a subsequent test phase, participants were asked to either rate some new robots as how likely they belong to the family of robots they were just shown, or build more robots that belong to the same family. Our model-based analysis showed that participants' generalization behavior in the test phase can be best explained by a library learning model that uses probabilistic programs as the underlying representations of the stimuli and exploits the regularities of the stimuli by extracting subroutines that are both simple and lead to compact representations of the input.

[1] Error bars reflect the relative contrast between different libraries across participants, not the uncertainty for the estimate.

Our work builds on the classic “analysis-by-synthesis” idea that sensory data can be more richly represented by modeling the process that generated it. Specifically, we model learning visual concepts as synthesizing structured, generative programs that generated these concepts. This approach has wide theoretical implications in many other domains. In particular, research has shown that handwriting may facilitate the acquisition of literacy in young children (James & Engelhardt, 2012), speech perception and language comprehension may involve access to the speech motor system (Liberman et al., 1967), visual production and recognition are functionally related in the brain (Fan et al., 2020), etc. In addition to the behavioral evidence, there has also been work showing that computational models that adopt the “analysis-by-synthesis” approach well captured human behavior: for example, in one-shot generalization of hand-written characters (Lake et al., 2015), learning morpho-phonology of human language (Ellis et al., 2022), perceiving 3D objects when certain visual features were missing Yildirim et al., 2024, etc. Our results from the current study are broadly in line with previous findings.

The current study focuses on how people acquire perceptual abstractions from the visual input. Specifically, participants were passively exposed to a stream of visual stimuli. One question left answered by this study concerns the role of action in abstraction learning. In this regard, in a similar line of work, McCarthy et al. (2023) found that people learn to solve physical assembly problems (i.e. building block towers) accurately and quickly across repeated attempts, and that the increase in accuracy reflects convergence on a small set of building policies among the group of participants. These physical assembly policies discovered are reminiscent of the subroutines (e.g. chunk subroutines in particular) learned by our library learning model in this paper, the key difference is that participants actively *constructed* the stimulus in each trial in the physical assembly experiments rather than mere observing. One key direction for future work is to examine to what degree *perceptual* abstractions acquired from passive observation are different from the *procedural* abstractions acquired from actively constructing. It is worth noting that the “analysis-by-synthesis” paradigm does not predict a clear difference between these two kinds of abstractions: because it always assumes a “synthesis” step being involved when acquiring new representations, regardless of the task being perceptual or procedural. Future experiment could be designed to investigate the differences between learning these two kinds of abstractions in terms of the learning speed, robustness to noise, etc. This will not only contribute to our understanding of how people learn abstractions across different tasks, but will also enrich the “analysis-by-synthesis” theoretical framework.

In sum, our paper shows that people are able to acquire higher-order structure from the statistical regularities in the spatial configurations of visual stimuli, and investigates how this learning is shaped by the distribution statistics of the visual stimuli. In the long run, such studies may lead to better understanding of the inherently generative and compositional organization of human conceptual knowledge (Palmer, 1977; Tenenbaum et al., 2011).

1.6 Acknowledgments

Chapter 1, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Wang, H., Polikarpova, N., & Fan, J. (2021). Learning part-based abstractions for visual object concepts. In *Proceedings of the 43th Annual Meeting of the Cognitive Science Society*. The dissertation author was the primary investigator and author of this material.

References

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review*, *8*(4), 629–647.

Ashby, F. G., Alfonso-Reese, L. A., Waldron, E. M., et al. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*(3), 442.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, *9*(4), 321–324.

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive psychology*, *63*(4), 173–209.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric bayesian framework for constructing flexible feature representations. *Psychological review*, *120*(4), 817.

Bliss, C. I. (1934). The method of probits. *Science*, *79*(2037), 38–39.

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38.

Bourne, L. (1974). An inference model for conceptual rule learning.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968.

Ellis, K., Albright, A., Solar-Lezama, A., Tenenbaum, J. B., & O'Donnell, T. J. (2022). Synthesizing theories of human language with bayesian program induction. *Nature communications*, *13*(1), 5024.

Ellis, K., Wong, L., Nye, M., Sable-Meyer, M., Cary, L., Anaya Pozo, L., Hewitt, L., Solar-Lezama, A., & Tenenbaum, J. B. (2023). Dreamcoder: Growing generalizable, interpretable knowledge with wake–sleep bayesian program learning. *Philosophical Transactions of the Royal Society A*, *381*(2251), 20220050.

Erickson, M. A. (2008). Executive attention and task switching in category learning: Evidence for stimulus-dependent representation. *Memory & Cognition*, *36*, 749–761.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107.

Fan, J. E., Wammes, J. D., Gunn, J. B., Yamins, D. L., Norman, K. A., & Turk-Browne, N. B. (2020). Relating visual production and recognition of objects in human visual cortex. *Journal of Neuroscience*, *40*(8), 1710–1721.

Ferguson, B., Franconeri, S. L., & Waxman, S. R. (2018). Very young infants learn abstract rules in the visual modality. *PloS one*, *13*(1), e0190185.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, *12*(6), 499–504.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman; Hall/CRC.

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In *Perceptual organization in vision* (pp. 245–290). Psychology Press.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in cognitive sciences*, *4*(5), 178–186.

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

James, K. H., & Engelhardt, L. (2012). The effects of handwriting experience on functional brain development in pre-literate children. *Trends in neuroscience and education*, *1*(1), 32–42.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133.

Katona, G. (1940). Organizing and memorizing: Studies in the psychology of learning and teaching.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 811.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

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

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Little, D. R., Nosofsky, R. M., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 1.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, *38*(4), 465–494.

McCarthy, W. P., Kirsh, D., & Fan, J. E. (2023). Consistency and variation in reasoning about physical assembly. *Cognitive Science*, *47*(12), e13397.

McCloud, S. (1994). *Understanding comics*. HarperPerennial. https://books.google.com/books?id=oJ1vPwAACAAJ

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. In *Psychology of learning and motivation* (pp. 207–250). Elsevier.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive psychology*, *9*(4), 441–474.

Plotkin, G. (1970). *Lattice theoretic properties of subsumption*. Edinburgh University, Department of Machine Intelligence; Perception. https://books.google.com/books?id=2p09cgAACAAJ

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3p1), 353.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition*, *81*(2), 149–169.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current directions in psychological science*, *12*(4), 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*(4), 606–621.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*(2), 273–284.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*(5306), 1599–1603.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Wang, H., Polikarpova, N., & Fan, J. E. (2021). Learning part-based abstractions for visual object concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How do infants become experts at native-speech perception? *Current Directions in Psychological Science*, *21*(4), 221–226.

Yildirim, I., Siegel, M. H., Soltani, A. A., Ray Chaudhuri, S., & Tenenbaum, J. B. (2024). Perception of 3d shape integrates intuitive physics and analysis-by-synthesis. *Nature Human Behaviour*, *8*(2), 320–335.

Chapter 2

Composing physical concepts to generate physical predictions

Abstract

The ability to make reliable predictions about the motion of objects is crucial for navigating the physical world. Here we investigate how people simultaneously infer multiple physical variables based on initial observations of how various objects behaved and spontaneously compose knowledge about these variables to make predictions in a novel scenario. Participants (N=203) observed objects launched at different angles in a 2D virtual environment and generated predictions about their trajectories. We found that participants' predictions were consistent with having accurately inferred each object's mass, as well as the effect of an invisible "wind" force that sometimes perturbed the motion of these objects. Critically, participants were also able to leverage these inferences to make accurate predictions even under conditions they had not previously experienced. Furthermore, we found that their generalization behavior was well explained by a computational model that embodied an accurate representation of the relevant physical variables (i.e., mass, wind), and better explained by this model than several plausible alternatives. Together, these findings provide insight into how people rapidly adapt their intuitive understanding of physical dynamics to make predictions in novel settings.

2.1 Introduction

A fundamental challenge humans face is to generalize from limited experience to new situations. This challenge manifests pervasively in how people understand and navigate the physical world. In particular, people need to be able to form expectations about how physical objects behave that are general enough to extend to settings they have not had direct experience with before. For example, a golfer who has not played on a windy day before might still need to anticipate how the wind will affect a golf ball's motion. Or a driver navigating snowy conditions for the first time might need to anticipate how the slippery roads will affect how their car needs to be handled. What cognitive mechanisms explain how people readily meet these challenges in everyday life?

One prominent account is that people's intuitive understanding of physical dynamics relies on having an internal model that simulates the motion of objects based on their physical properties — or in other words, an "Intuitive Physics Engine" (IPE) (Battaglia et al., 2013; Ullman et al., 2017). In particular, prior work in intuitive physics has established that people make accurate physical predictions about object motion that are consistent with the predictions of an IPE applying approximately Newtonian mechanics under uncertainty (Smith, Battaglia, et al., 2013; Smith, Dechter, et al., 2013; Smith & Vul, 2013, 2015). Moreover, it has been shown that people can leverage their observations of how an object moves to infer latent physical properties of objects, such as their mass, gravity, friction, elasticity, and deformability (Bramley et al., 2018; Hamrick et al., 2016; Sanborn et al., 2013; Tung et al., 2023; Ullman et al., 2018). However, a key property of these IPE models is their ability to flexibly combine inferences about multiple such properties to generate predictions in new situations (Battaglia et al., 2013; Kubricht et al., 2017; Lake et al., 2017). While recent work in intuitive physics has investigated transfer of a single physical property (e.g. mass) inferred in one task setting to another (Allen, Smith, et al., 2020; Neupärtl et al., 2020), it remains unclear how people carry inferences about multiple physical properties (e.g., mass and friction) into new environments and to what degree their patterns of generalization are well explained by the IPE hypothesis.

Furthermore, a key question concerns the precision with which objects and forces are mentally represented to enable such generalization. For example, on a day when there is a gust of wind blowing towards the northeast at 16mph, how do golfers represent what forces are operative in order to make good predictions? Do they represent gravity (which drags objects downward) and the wind (which is pointing to the northeast) as two separate forces? Or do they effectively treat these two forces as one net force pointing towards the northeast but slightly downward? While the former is more accurate, the latter is simpler while still leading to approximately correct predictions in many cases. Indeed, recent work has found that modeling intuitive physical judgments using simpler representations (e.g., approximations of an object's shape) can sometimes align better with human judgments than using more faithful representations of physical reality (Li et al., 2023). How could we tell what kinds of representational commitments people implicitly make when generating predictions about physical dynamics? Previous studies have generally investigated this question by eliciting categorical judgments about each latent physical property in question (e.g., *Which ball is heavier? Is there a lot of friction in this region?*) (Hamrick et al., 2016; Ullman et al., 2018). However, this approach offers limited resolution concerning people's estimates of these properties, and thus is not well suited for disentangling hypotheses concerning how precisely they are represented.

The current study builds on prior work to investigate the mental representations for objects and forces people use to generalize to novel settings. Participants first learned about multiple physical properties (i.e., the masses of several objects, the forces in different environments) and were later challenged to combine aspects of what they had to learn in novel ways. We also leverage continuous report measures to obtain more precise estimates of both participant's predictions concerning the observable scene dynamics, but also latent physical properties of the objects within these scenes, consistent with prior work in intuitive

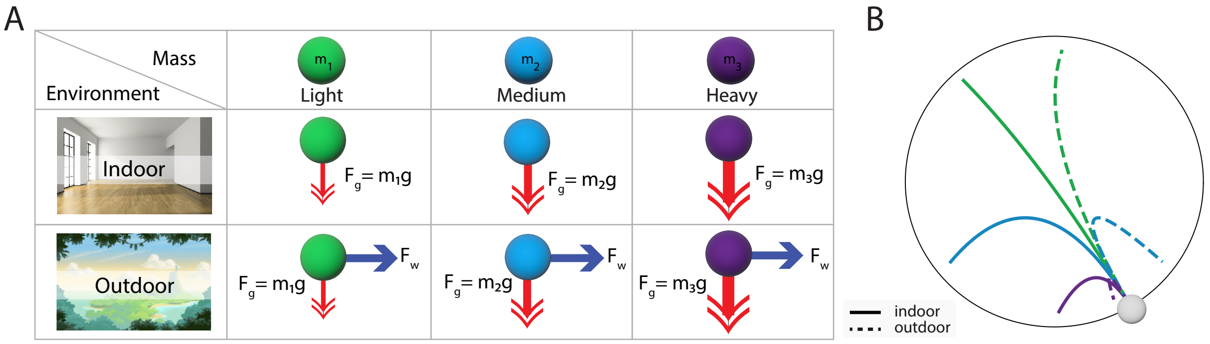

Figure 2.1. (A) The 2 × 3 design matrix of our experiment, where participants were trained on 5 out of these 6 cells, and asked to generalize to the held-out cell. The choice of held-out cell was counterbalanced across participants. (B) Different trajectories of a ball when its mass and the environment varies.

physics that has used similar continuous report measures (Allen, Smith, et al., 2020; Allen, Smith, et al., 2020; Dasgupta et al., 2018; Smith & Vul, 2013, 2015; Zhou et al., 2023). Finally, we define several computational cognitive models that embody different hypotheses concerning the contents of the mental representation people use to represent objects and forces, enabling quantitative comparison between models that are more faithful to physical reality and others that use simpler approximations for explaining human judgments.

2.2 Methods

2.2.1 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 institution's IRB. The experiment lasted approximately 35 minutes and participants were paid $14/hr based on this expected completion time.

2.2.2 Task environment & procedure

In this experiment, participants played a virtual game of catch. On each trial, participants used the arrow keys to move a rectangular paddle along the outside of the circle to intercept a ball launched towards the center from another location on the circle (Figure 2.1B), thereby providing a continuous measure of where on the circle participants expected the ball to land. Once participants were satisfied with the paddle's location, they pressed the space bar to launch the ball and observe the actual trajectory it took. If the ball made contact with any part of the paddle, this was considered a success. The paddle always began in the same starting location (i.e., 3 o'clock), and thus its starting location was not predictive of where the ball would land. Across trials, the ball was launched with either a larger or smaller amount of force, the magnitude of which was indicated by the length and thickness of an arrow originating at the ball's starting location and pointing towards the center of the circle.

Given the observed starting location and launching force, the ball's motion on each trial depended on two additional latent factors: the mass of the ball and the forces in the environment. There were three balls: 'light', 'medium' and 'heavy.' The balls were of the same size but distinguished by color and texture cues, allowing participants to learn across trials to associate each ball's visual properties with how it behaved. Which of these balls was light, medium, and heavy was randomized across participants. There were two environments with different forces: an 'indoor' and 'outdoor' environment. Each of these environments was cued by an evocative background image: an indoor scene image for the indoor environment and outdoor scene image for the outdoor environment. In the indoor environment, the only force was a downward gravitational force (F_g); in the outdoor environment, there was both a downward gravitational force and a rightward wind-like force (F_w).

The experimental session was divided into two phases: a training phase and a test phase. During the training phase, participants completed four blocks of trials, such that in each block people repeatedly made predictions about multiple balls in one of the environments. Critically, however, participants were only exposed to five of the six possible combinations of ball masses and environments. The sequence of training blocks was such that exposure to each environment was temporally counterbalanced (i.e., ABBA) within participant, and which environment participants encountered first was also counterbalanced across participants. Within each block, participants observed the ball launched from all 12 possible launching locations uniformly spanning the full circumference of the circle, and with both launching forces (i.e., larger, smaller) at each of these locations. In total, the training phase consisted of 120 trials: 12 starting locations × 2 launching forces × 5 mass-environment combinations. Transitions between each blocks were not marked in any way.

During the test phase, participants performed the same task but with a novel ball-mass and environment combination, thus requiring them to compose what they had inferred about these latent factors in a new way. Which mass-environment combination appeared in the test phase was counterbalanced across participants.

2.3 Results

2.3.1 People learn the dynamics

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 2.2A 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 2.2B 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, Figure 2.2B 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.

2.3.2 People learn and combine force and mass

On average, test-phase behavior exhibits correct force-mass combination

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

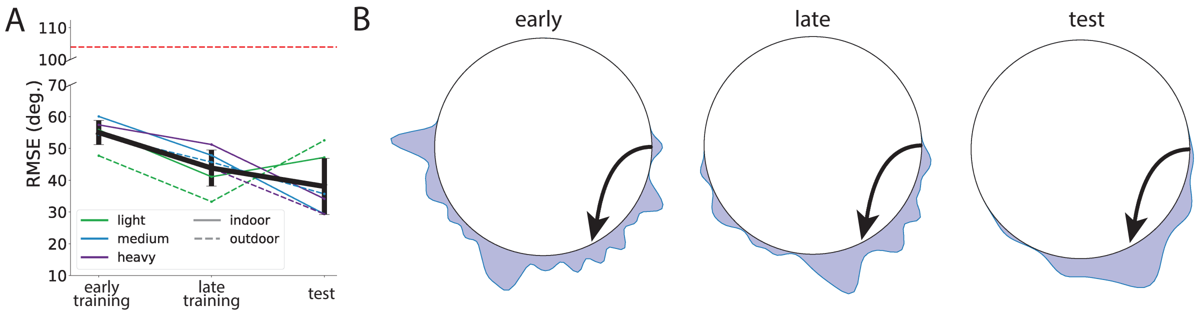

Figure 2.2. (A) RMSE for all 6 conditions, black thick line shows the mean and standard deviation. Dashed red line represents expected performance under random guessing (Rinaman et al., 1996). (B) Three trials with the same launching condition were selected from the three timepoints. Black-dotted trajectory 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).

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 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:

$$P(M|D) = \frac{P(D|M)P(M)}{\sum_{M'} P(D|M')P(M')} \tag{2.1}$$

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

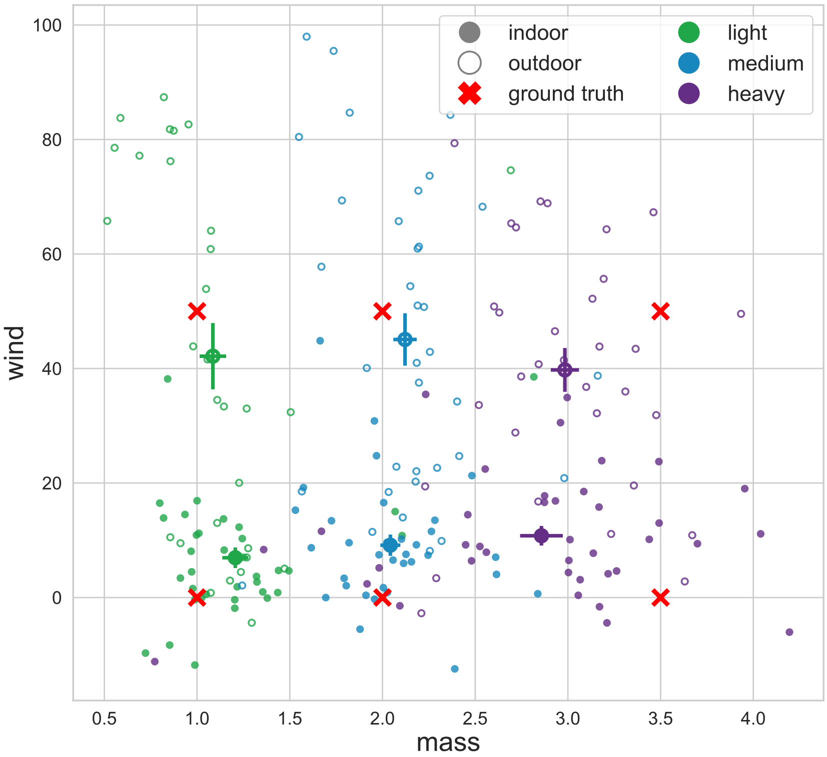

Figure 2.3. Fitted parameters for each individual for all conditions, wind (F_w) on the y axis and mass (m) on the x axis. Error bars represent 95% CIs on each dimension. Red crosses indicates the ground truth wind and mass values for the 6 conditions.

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 2.2B 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 2.3 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).

Individuals correctly combine the force and mass they learned

Our model-based analysis revealed substantial variation between individuals (Figure 2.3), 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

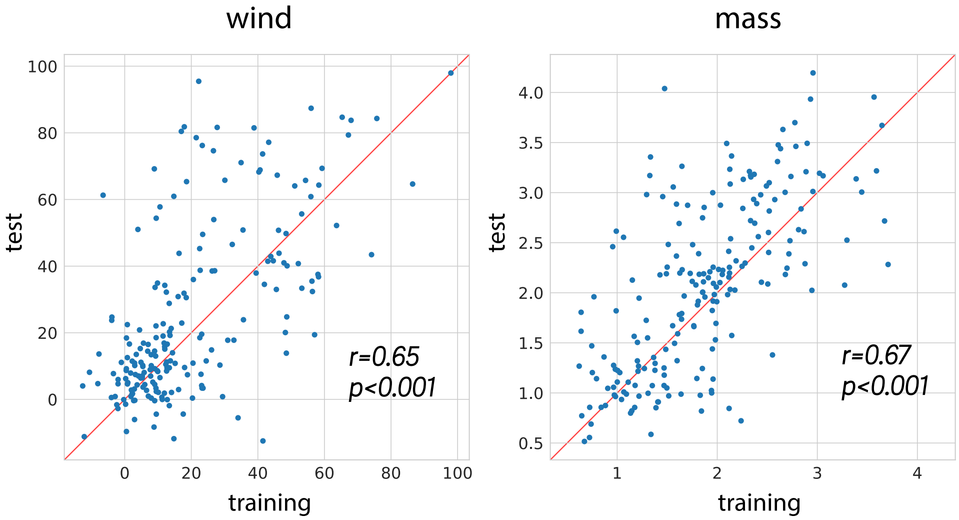

Figure 2.4. Fitted parameters (wind w and mass m) for each individual for training and test, training on the x axis and test on the y axis. Red line is the diagnal line.

(even if they are not veridical)?

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 2.1, 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 2.4), 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

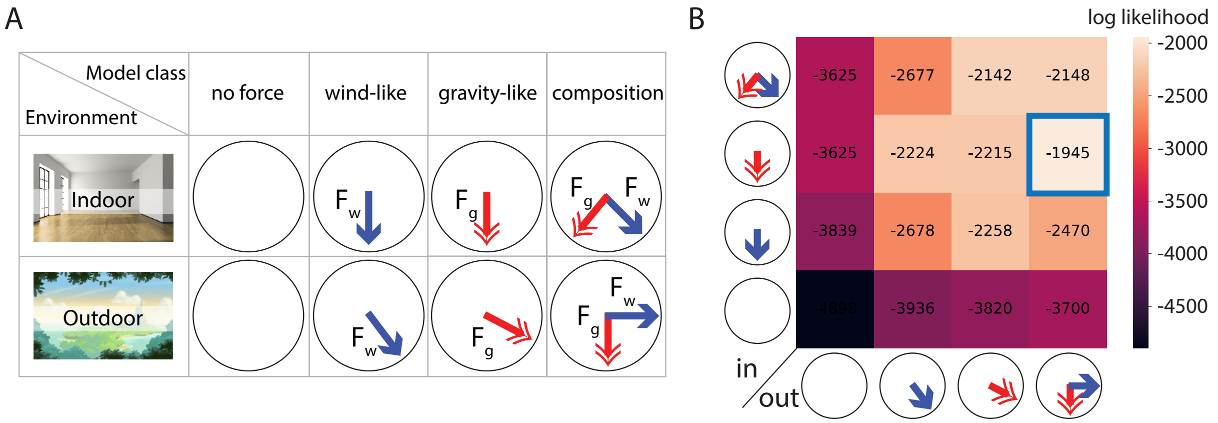

Figure 2.5. (A) An illustration of the 4 types of models we considered in each environment. (B) log-likelihoods of 16 different models, the model with the highest log-likelihood is colored in blue boarder.

and test phases, even when their estimate of either a ball's mass or the wind was not veridical.

2.3.3 Human behavior is most consistent with the ground-truth force structure.

In the last section, we showed that when assuming participants are using the correct world model, namely, there is a downward gravity in the indoor environment, a downward gravity and a rightward wind in the outdoor environment, their fitted parameters can recover the ground truth wind and mass values. However, is it possible that participants' behavior can be better explained by having other world models? For example, in the outdoor environment, rather than postulating that there is wind and gravity, is it possible that participants have a simpler model in mind, thinking that there is just one force pointing towards the bottom-right corner? This is a legitimate concern and cannot be ruled out using the previous analysis.

Furthermore, we only considered motor noise while fitting the parameters in the last section, prior work (Battaglia et al., 2013; Li et al., 2023; Smith & Vul, 2013), however, has shown that people are subject to a wide range of sources of uncertainty when making physical predictions. How, then, will the conclusion from the previous section change when considered in this more realistic setting? We aim to address these two concerns in this section.

To evaluate whether participants' behavior is more consistent with an alternate world model, we fit 16 different classes of models with different force structures to the observed ball trajectories, and ask whether participants' behavior is consistent with the predictions of these models under any combination of perceptual and motor noise.

Specifically, we considered two kinds of force structures, one is gravity-like, meaning that the force is proportional to the mass of objects; and another is wind-like constant force. This gives us four possible world models in each environment: no force, wind-like force only, gravity-like force only and the composition of gravity-like and wind-like forces (see Figure 2.5A). Considering that participants may use different world models for the indoor and outdoor environments, we consider all 16 possible combinations of world models in total (see Figure 2.5B). We take the "best performing variant" of each class by optimizing their free parameters (the mass of the three balls, the magnitude and direction of the constant and proportional-to-mass force) using the ball's ground truth trajectories. An example of the parameters for each of the models after optimization are shown in Figure 2.5A.

In order to see which model is participants' behavior most consistent with, we calculate the probability of the participants' data under each model's prediction distribution. We consider three kinds of perceptual uncertainty in the initial launching condition, from a wide range of values: uncertainty over the launching force (σ_F, in ranges [1, 25]), uncertainty over the launching angle (σ_θ, in ranges [0.1 rad, 2 rad]) and uncertainty over the launching position (σ_ρ, in ranges [0.04 rad, 0.4 rad]). We also consider participants' motor noise (σ_m, in ranges [0.1 rad, 0.5 rad]) when placing the paddle given their prediction (see Figure 2.6 for details of how each of these noise parameters is defined).

The likelihood of each model is calculated by marginalizing over all noise parameter

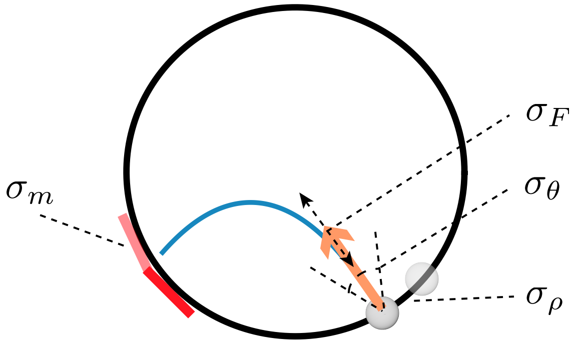

Figure 2.6. Illustration of different kinds of uncertainty we considered: uncertainty over the launching force (σ_F), uncertainty over the ball's launching angle (σ_θ), uncertainty over the ball's launching position (σ_ρ), and motor noise (σ_m) when placing the paddle.

combinations, defined as follows:

$$P(D|M) = \int_{\boldsymbol{\theta}} P(D|M, \boldsymbol{\theta}) P(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta} \tag{2.2}$$

where D is participants' paddle placements during generalization, M is one of the 16 different models, $\boldsymbol{\theta}$ refers to $\{\sigma_F, \sigma_\rho, \sigma_\theta, \sigma_m\}$. We use half of the participants' data to obtain $P(\boldsymbol{\theta}|M)$ (posterior of $\boldsymbol{\theta}$ under the data using a uniform prior), and use the other half participants' data to score the models. We do so for 100 random splits and calculate the average of each of the model's log-likelihoods, shown in Figure 2.5 B.

We found that the log-likelihood of the ground truth model (i.e. gravity-like force in the indoor environment and composition of gravity-like and wind-like forces in the outdoor environment) is higher than all the alternative models, this implies that not only is participants' generalization behavior consistent with ground-truth parameters in a model with ground-truth force structure, but their behavior is also most consistent with this ground-truth model than any other plausible force-structure model, even when multiple sources of uncertainty are considered.

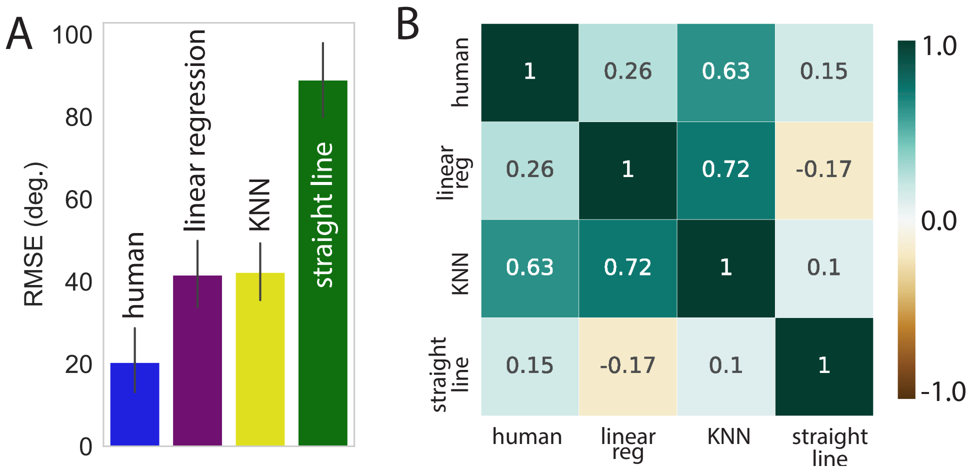

Figure 2.7. (A) RMSE of average response for models and human, error bars reflect 95% CIs using 10,000 iterations bootstrapping. (B) Correlation between different model and human's signed errors.

2.3.4 Comparing different heuristic 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. In this section, we explore several other 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 heuristic 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.

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. 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 2.7A shows RMSE of the models compared to humans. The straight line heuristic

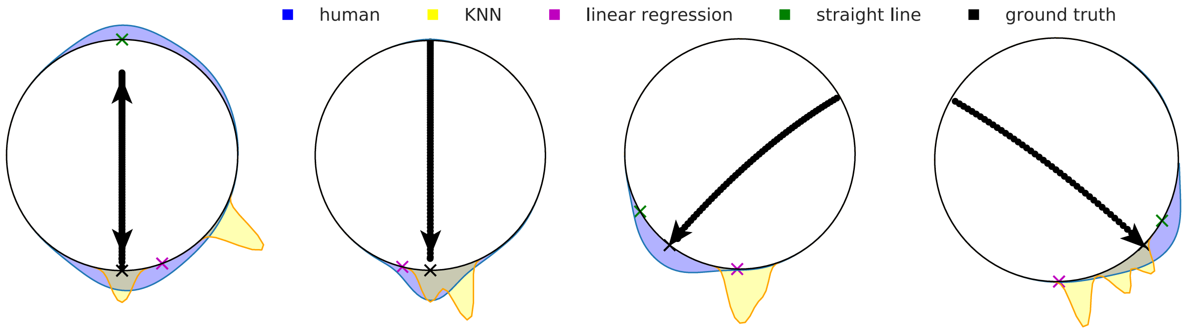

Figure 2.8. Four trials selected from the test phase. In these trials, the ball was launched from 6 o'clock, 12 o'clock, 2 o'clock and 10 o'clock in the indoor environment respectively. KNN and human predictions are shown in distribution. Straight line model and linear regression model's predictions are shown by the × marker because they are deterministic (see text). We also show the ground truth trajectory and landing location of the ball in each plot.

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. We further calculated the correlations between each model and human's signed errors, as shown in Figure 2.7B, 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. We can see that none of the alternative models has a high correlation with participants' behaviors, suggesting that none of these alternative models provides a satisfying account of how humans were able to perform this task.

2.4 Discussion

What cognitive mechanisms enable people to leverage prior experience to quickly calibrate their physical predictions in new environments? To address this question, we conducted a study where participants viewed different objects being launched under simulated physics and had to predict their trajectory. During the initial learning phase, they observed each object in at least one of two environments, wherein different ambient forces were present (i.e., one was "windy" while the other was not). In a subsequent test phase, they had to predict how one of these objects would behave in an environment other than the one where they had initially encountered it, thus requiring them to generalize what they had learned about that object and environment to make accurate predictions. Participants' response patterns suggest that their predictions in this test phase reflected accurate inferences about the masses of each object and the forces operating in each environment, rather than reliance upon simple heuristics (e.g., approximating the joint contribution of "gravity" and "wind" into a single net force). More generally, these findings are consistent with the notion that people can readily combine aspects of what they have learned about objects and forces in novel ways to rapidly generalize to new task conditions.

Our findings contribute to a broader literature that has investigated how people learn generalizable rules and patterns that go beyond encoding specific associations between individual observations. In particular, we showed that participants were able to distinguish forces with different *functional forms* (e.g., wind exerts a constant force whereas the strength of gravity is a function of mass). These findings are reminiscent of those from prior studies of function learning, where participants learn to map values of one variables to values of another. It has been found that people can learn and recall a wide variety of relationships between continuous variables, including linear functions, higher-degree polynomial functions, and power-law relationships (Brehmer, 1971, 1974; Busemeyer et al., 2013; Carroll, 1963; DeLosh et al., 1997; Kalish et al., 2004; Koh & Meyer, 1991; Mcdaniel

& Busemeyer, 2005). Similar ideas have also been explored in studies of reward learning (Choung et al., 2017; Schulz et al., 2018; Song et al., 2022; Speekenbrink, 2022; Wu et al., 2018), where participants learn to map their own decisions in systematic ways to predict rewards, as well as category learning (Ashby & Maddox, 2005; Ballard et al., 2018; Mack et al., 2016), where people aim to extract abstract rules that are diagnostic of members of a given category. Taken together, these findings are compatible with the possibility that similar cognitive mechanisms might support learning of abstract functional relationships between variables across a wide range of behavioral domains (Kemp & Tenenbaum, 2008).

Our findings also have potential implications for advancing computational theories of human-like intuitive understanding of physical scene dynamics from visual inputs. In particular, recent studies suggest that there is still a significant performance gap between current vision models and humans (Bear et al., 2021; Buschoff et al., 2023; Duan et al., 2022; Tung et al., 2023). While it is not entirely unclear what ingredients are needed to develop vision models that better approximate human physical scene understanding, our findings are consistent with existing proposals suggesting that ability to extract more explicit structured representations of objects, their attributes, and forces operating on them could be especially crucial (Bear et al., 2021).

Because the current study focused on generalization in the test phase, an important question left open concerns the *dynamics* by which people updated their mental representation of the objects and forces in each environment throughout the learning phase. One possibility is that participants might have serially tested different qualitative hypotheses about these entities, perhaps punctuated by flashes of insight (Gong & Bramley, 2023; Gong et al., 2023; Rule et al., 2020; Ullman et al., 2012; Ullman et al., 2018; Wang et al., 2021; Zhao et al., 2024). Alternatively, the learning we observed might have been driven primarily by slow error-driven learning mechanisms (Benson et al., 2011; Bond & Taylor, 2017; Hegele & Heuer, 2010; Heuer & Hegele, 2009; McDougle et al., 2015; Morehead et al., 2015; Taylor et al., 2014). Our exploratory analyses of learning curves at the group level are in principle compatible with both underlying smooth trend and abrupt transitions at the individual level. Towards distinguishing these possibilities, one valuable direction for future work might be to obtain denser measurements of learning dynamics at the individual level that enable quantitative modeling of these individual trajectories.

One limitation of our study as it pertains to real-world physical prediction is the restriction to 2D environments. In addition, the stimuli used in the current study are highly simplified: for example, the objects always traveled in parabolas and there were no collisions between objects. One promising avenue for future work might be to extend the current studies to more thoroughly investigate generalization across a wider variety of more complex 3D environments (Bear et al., 2021; Martinez et al., 2023; Tung et al., 2023).

Another key direction for future work is to examine what role prior knowledge plays in the process of inferring latent physical properties. In the current study, "gravity" always drags objects downwards and "wind" blows in a horizontal direction. Because these are familiar orientations for these types of forces, participants might have been able to more strongly rely upon prior knowledge to learn the properties of each object and environment in our study. Future investigations could introduce more unusual configurations of forces that might require participants to acquire internal representations of scene dynamics that depart more substantially from those they have previously encountered. Such studies would be helpful for understanding how much evidence is needed to adapt to these more unusual environments, and what cognitive mechanisms are recruited to enable such adaptation.

In summary, our paper provides an account of how people leverage prior experience to update their mental model of physical objects and forces, enabling them to compose aspects of this mental model in new ways to generalize quickly to new task demands. In the long term, such studies might contribute to more unified theories of the inductive biases and learning mechanisms that enable rapid learning and flexible generalization in humans.

2.5 Acknowledgments

Chapter 2, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Wang, H., Allen, K., Vul, E., & Fan, J. (2022). Generalizing physical prediction by composing forces and objects. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.* The dissertation author was the primary investigator and author of this material.

References

Allen, K. R., Smith, K., Piterbarg, U., Chen, R., & Tenenbaum, J. (2020). Abstract strategy learning underlies flexible transfer in physical problem solving. *CogSci*.

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, *56*, 149–178.

Ballard, I., Miller, E. M., Piantadosi, S. T., Goodman, N. D., & McClure, S. M. (2018). Beyond reward prediction errors: Human striatum updates rule values during learning. *Cerebral Cortex*, *28*(11), 3965–3975.

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., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., et al. (2021). Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

Benson, B. L., Anguera, J. A., & Seidler, R. D. (2011). A spatial explicit strategy reduces error but interferes with sensorimotor adaptation. *Journal of neurophysiology*, *105*(6), 2843–2851.

Bond, K. M., & Taylor, J. A. (2017). Structural learning in a visuomotor adaptation task is explicitly accessible. *Eneuro*, *4*(4).

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive psychology*, *105*, 9–38.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science*, *24*(6), 259–260.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.

Buschoff, L. M. S., Akata, E., Bethge, M., & Schulz, E. (2023). Have we built machines that think like people? *arXiv preprint arXiv:2311.16093*.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (2013). Learning functional relations based on experience with input–output pairs by humans and artificial neural networks. In *Knowledge concepts and categories* (pp. 405–437). Psychology Press.

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*(2), i–144.

Choung, O.-h., Lee, S. W., & Jeong, Y. (2017). Exploring feature dimensions to learn a new policy in an uninformed reinforcement learning task. *Scientific reports*, *7*(1), 17676.

Dasgupta, I., Smith, K. A., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2018). Learning to act by integrating mental simulations and physical experiments. *BioRxiv*, 321497.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968.

Duan, J., Dasgupta, A., Fischer, J., & Tan, C. (2022). A survey on machine learning approaches for modelling intuitive physics. *arXiv preprint arXiv:2202.06481*.

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

Gong, T., & Bramley, N. R. (2023). Continuous time causal structure induction with prevention and generation. *Cognition*, *240*, 105530.

Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N. R. (2023). Active causal structure learning in continuous time. *Cognitive Psychology*, *140*, 101542.

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.

Hegele, M., & Heuer, H. (2010). Implicit and explicit components of dual adaptation to visuomotor rotations. *Consciousness and cognition*, *19*(4), 906–917.

Heuer, H., & Hegele, M. (2009). Adjustment to a complex visuo-motor transformation at early and late working age. *Ergonomics*, *52*(9), 1039–1054.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological review*, *111*(4), 1072.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 811.

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Li, Y., Wang, Y., Boger, T., Smith, K. A., Gershman, S. J., & Ullman, T. D. (2023). An approximate representation of objects underlies physical reasoning. *Journal of Experimental Psychology: General*.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208.

Martinez, J., Binder, F., Wang, H., Haber, N., Fan, J., & Yamins, D. L. (2023). Measuring and modeling physical intrinsic motivation. *arXiv preprint arXiv:2305.13452*.

Mcdaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic bulletin & review*, *12*(1), 24–42.

McDougle, S. D., Bond, K. M., & Taylor, J. A. (2015). Explicit and implicit processes constitute the fast and slow processes of sensorimotor learning. *Journal of Neuroscience*, *35*(26), 9568–9579.

Morehead, J. R., Qasim, S. E., Crossley, M. J., & Ivry, R. (2015). Savings upon re-aiming in visuomotor adaptation. *Journal of neuroscience*, *35*(42), 14386–14396.

Neupärtl, N., Tatai, F., & Rothkopf, C. A. (2020). Intuitive physical reasoning about objects' masses transfers to a visuomotor decision task consistent with newtonian physics. *PLoS Computational Biology*, *16*(10), e1007730.

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

Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in cognitive sciences*, *24*(11), 900–915.

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

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of experimental psychology: learning, memory, and cognition*, *44*(6), 927.

Smith, K. A., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. *Proceedings of the annual meeting of the cognitive science society*, *35*(35).

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over time. *Proceedings of the annual meeting of the cognitive science society*, *35*(35).

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Smith, K. A., & Vul, E. (2015). Prospective uncertainty: The range of possible futures in physical prediction. *CogSci*.

Song, M., Baah, P. A., Cai, M. B., & Niv, Y. (2022). Humans combine value learning and hypothesis testing strategically in multi-dimensional probabilistic reward learning. *PLoS computational biology*, *18*(11), e1010699.

Speekenbrink, M. (2022). Chasing unknown bandits: Uncertainty guidance in learning and decision making. *Current Directions in Psychological Science*, *31*(5), 419–427.

Taylor, J. A., Krakauer, J. W., & Ivry, R. B. (2014). Explicit and implicit contributions to learning in a sensorimotor adaptation task. *Journal of Neuroscience*, *34*(8), 3023–3032.

Tung, H.-Y., Ding, M., Chen, Z., Bear, D., Gan, C., Tenenbaum, J. B., Yamins, D. L., Fan, J. E., & Smith, K. A. (2023). Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *arXiv preprint arXiv:2306.15668*.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

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.

Wang, H., Vul, E., Polikarpova, N., & Fan, J. E. (2021). Theory acquisition as constraint-based program synthesis. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, *2*(12), 915–924.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2024). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, *8*(1), 125–136.

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023). Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*.

Chapter 3

Generalizing physical predictions across various scenarios

Abstract

How do people perform general-purpose physical reasoning across a variety of scenarios in everyday life? Across two studies with seven different physical scenarios, we asked participants to predict whether or where two objects will make contact. People achieved high accuracy and were highly consistent with each other in their predictions. We hypothesize that this robust generalization is a consequence of being able to run noisy physics mental simulations. We designed an "intuitive physics engine" model to capture this generalizable simulation. We find that this model generalized in human-like ways to unseen stimuli and to a different query of predictions. We further evaluated several state-of-the-art deep learning and heuristics models on the same task and found that they could not explain human predictions as well. This study provides evidence that human's robust generalization in physics predictions are supported by a probabilistic simulation model, and suggests the need for structure in learned dynamics models.

3.1 Introduction

Every day, we interact with the physical world in a variety of ways. We might start the morning by pouring cereal into a bowl, and later stably stacking that bowl with the rest of the dishes in the sink. Around the house, we might use a book to stabilize a wobbly chair, or throw trash into the trash can. Later with our kids we might build towers with blocks, or determine the right shot in a game of billiards. All of these scenarios require knowledge of many different principles of physics: from containment to stability to ballistic motion to collision dynamics. Yet we handle each of these tasks naturally, and often with little effort. But how are people able to do such general-purpose physical reasoning?

One hypothesis that has grown in prominence over the past decade is that we have a cognitive module that can perform general purpose, probabilistic physics simulation, often termed the *Intuitive Physics Engine* (Battaglia et al., 2013; K. Smith et al., 2024; Ullman et al., 2017). Under this hypothesis, general physics understanding arises because the simulation engine contains more primitive components for modeling the world – representations of objects and the forces they exert on each other, latent properties such as mass or elasticity that constrain how objects respond to forces, and key dynamic quantities such as momentum and events such as collisions – and combines them with uncertainty about the state of the world to reason probabilistically about a wide range of scenarios we might expect to encounter in day-to-day life. Thus, much like the physics engines that underlie many computer simulations, these building blocks of knowledge can be combined to model much more complex and combinatorial situations.

While this hypothesis has received quantitative support from many studies, a crucial aspect of it has never been explicitly tested. Prior studies that model human intuitive physics have typically focused on just one scenario at a time: e.g., how or whether a stack of objects might fall (Battaglia et al., 2013; Hamrick et al., 2016; Zhou et al., 2023), how moving objects will bounce off each other and fixed obstacles (Gerstenberg et al., 2021;

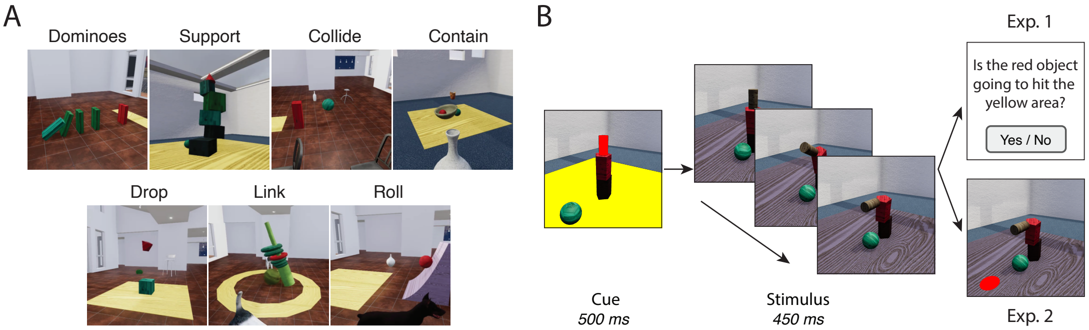

Figure 3.1. A: The seven different scenarios testing different physics principles. In each image, the object colored in red is the target object and the yellow area on the ground is the zone. B: Participants are cued with the target and zone objects, observe a short video, and predict either whether or where the target will contact the zone.

Neupärtl et al., 2021; K. A. Smith & Vul, 2013; Ullman et al., 2018), or how liquid will pour (Bates et al., 2019; Kubricht et al., 2016). While these models all use a physics engine at their core, across this research, modelers make different assumptions about the particulars of the physics engine, and fit different parameters to capture uncertainty about the state of the scene or how physical events resolve. This modeling approach risks overfitting to specific scenarios, and thus cannot answer the question of whether people have a general purpose physics simulator, or use different systems for different physical principles.

Indeed, another theory of human physical reasoning is that our judgments are based on inferences from past experience. This idea was first manifested in exemplar-based models and simple heuristics (Gilden & Proffitt, 1989; Nusseck et al., 2007; Proffitt et al., 1990; Sanborn et al., 2013): that people might base their judgments exclusively on combinations of features of the initial scene configuration without explicit reference to physical dynamics. Similar idea has also been expressed by recent neural network models that learn to predict dynamics by watching videos. Proponents of this approach suggest that learning physics from raw data provides two benefits: these models can extract generalizable physical principles more flexibly than if the models were to rely on a fixed simulator, and can work directly from visual inputs in a way that physical simulation models on their own do not.

A range of models have been proposed that express a spectrum of assumptions about what parts of physics should be learned, from those that attempt to jointly learn a scene representation and dynamics with few assumptions about the structure (Babaeizadeh et al., 2020), to models that assume the scene structure is known and try to learn only how objects interact (Han et al., 2022), and many in between. While these neural networks are often intended purely to advance an AI system's understanding of the physical world, they have been proposed as hypotheses for how infants learn physics (Piloto et al., 2022), and have been used to predict both behavior and neural activity in monkeys performing physics prediciton tasks (Nayebi et al., 2023).

In this paper, we test whether human physical predictions can be explained by approximate probabilistic inference in a single, general physics simulator across a wide range of everyday settings. We use an adapted version of the Physion dataset (Bear et al., 2021) which was designed to test arbitrary models' physics understanding in a variety of scenarios against both ground truth and human beliefs. We specifically test the *generalizability* of models: how well models can explain human predictions in scenarios that they have not been fitted or trained on. We show that an intuitive physics engine generalizes to these unseen scenarios in a human-like way, explaining human behavior only slightly worse than expected by the noise ceiling. We compare a variety of state-of-the-art deep learning models that encompass a range of assumptions about what is learned. Some jointly learn representations and dynamics with little structure, testing whether physics can be learned directly from video. Others learn to parse images into scene representations in a variety of ways – from few assumptions about scene structure to strong assumptions about 3D world structure – and learn dynamics on top of that representation with a recurrent network, in order to test how well these models produce representations that support learning physics. We also compare multiple models that make physical predictions based on initial scene features. We find that an intuitive physics engine model captures human judgments and generalizes to unseen scenarios and novel tasks remarkably well, and far better than the deep learning and feature-based models tested. These results both support the mental simulation hypothesis as a generalizable mechanism for intuitive physical reasoning and point to the value of including stronger and more structured inductive biases into neural network models of intuitive physics.

3.2 Human experiments

To evaluate people's physical predictions across a wide range of scenarios, we adapted seven rigid body scenes from the Physion (Bear et al., 2021) dataset. These seven scenarios test a variety of physical concepts (Fig. 3.1A): chains of collisions (*dominoes*), the stability of a stack of objects (*support*), the particular ways collisions resolve (*collide*), whether one object can contain another (*contain*), how individual objects fall (*drop*), how a collection of objects can be knocked over (*link*), and how objects roll or slide down a slope (*roll*).

In each trial, there is a target object and zone. The goal is to predict either whether (Exp. 1) or where (Exp. 2) the target object will contact the zone at some point in the future.

Each scenario consists of 150 trials (1050 total), varying in scenario-specific configurations (see Bear et al. (2021) for details on their construction). Each trial consisted of a 450ms video in which the target object does not yet touch the zone. The trials were designed so that if the video had continued, in half of them the target would touch the zone within the next 2 seconds, but would never touch in the other half.

3.2.1 Experiment 1: Will it collide?

To first understand how well people make these physical predictions, we ask participants to make a binary judgement of whether they think the target object will contact the zone after watching a short video clip.

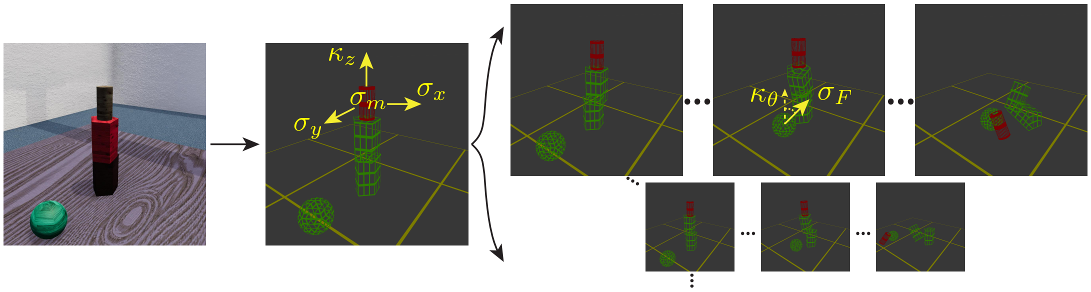

Figure 3.2. The Intuitive Physics Engine (IPE). The scene (left) is perceived to form an internal representation (middle), that includes perceptual (σ_x, σ_y, κ_z) and physical property (σ_m) uncertainty. The IPE then uses this representation to probabilistically simulate (right) how the world might unfold based on dynamic uncertainty (σ_F, κ_θ).

Participants

350 participants (50 per scenario; 198 female; all native English speakers) recruited from Prolific completed the experiment. Each participant was shown all 150 stimuli from a single scenario. Data from 33 participants were excluded for failing our preregistered inclusion attention checks. The experiment lasted approximately 15 minutes and participants were paid $3.50.

Task procedure

The structure of our task is shown in Fig. 3.1B. Each trial began with a 500-1500ms fixation cross. Participants then saw the first frame of the video for 500ms with the target object and zone flashing a red and yellow overlay respectively, followed by the stimulus video for 450ms. Participants then saw a screen with buttons to indicate "YES" (the target would contact the zone) or "NO" (it would not). Before the main task, participants observed 10 familiarization trials for which the full movie was shown post-prediction.

Results

We first examined how often participants' predictions of contact agreed with the simulation outcome from the physics engine used to create the stimuli. We found that people achieved high accuracy (proportion correct = 0.81, 95% CI=[0.77, 0.84]) and their performance was substantially above chance across all seven scenarios ($t(6)$ =16.49, $p < 10^{-5}$). Participants' data also demonstrated variation in performance across scenarios, achieving the highest accuracy in Dominoes (0.84) and lowest in Roll (0.74).

Even though people made errors on some trials, these errors were consistent across participants (cross-trial bootstrapped split-half reliability=0.94, 95% CI=[0.92, 0.96], Fig. 3.3A). This pattern of high but imperfect accuracy and reliable errors is especially useful when comparing models with humans: to be a good explanation of how people make physical predictions across scenarios, a model should not only achieve high accuracy, but also err in the same ways as people do.

3.2.2 Experiment 2: Where will they touch?

Here we investigate more fine grained predictions by asking participants to indicate where they believe the target object will first contact the zone.

Participants

A separate group of 245 participants (35 per scenario; 157 female; all native English speakers) recruited from Prolific completed the experiment. The experiment lasted approximately 16 minutes and paid \$3.75.

Task procedure

The task procedure was identical to the "Will it collide" task except that participants were asked to place a circular disk where they believed the target would first contact the zone (Fig. 3.1B). For each scenario, the stimuli were the same as the previous experiment except that we filtered the 150 trials to only include trials where (a) the target object contacted the zone, and (b) this collision happened at a location that was unoccluded by other objects (Collide: 50 trials, Contain: 44, Dominoes: 68, Drop: 63, Link: 63, Roll: 65, Support: 48). After showing the first 450ms of the stimulus, the video froze on the final frame and participants used their cursor to position a disk on the target zone. Only the part of the disk overlapping the zone was displayed. When participants had placed the disk at their desired location, they clicked a "NEXT" button to register their prediction.

Results

For each trial, we measured the center point of participants' disk placement positions as the 3D location in world coordinates. We excluded participants' placements that were off the zone by the disk radius (i.e. the disk had no overlap with the zone at all during the experiment, indicating that participants were not following the instructions or misclicked), accounting for about 5% of the data.

To assess how far off people are from the contact points given by the ground truth stimulus, we first calculated the Euclidean distance between the mean human predictions and the ground truth contact point for each trial, and averaged across trials. Because this metric is sensitive to the area of the target zone, for each trial we divided the distance by the standard deviation of participants' placements on that trial (analogous to $d^{'}$ in signal detection theory). We found similar patterns to the "will it" task: participants' predictions are significantly closer to ground truth than expected by chance, and this is true for every physical scenario (mean normalized distance=1.39, 95% CI=[1.23, 1.51], $t(6) = 43.25$, $p < 10^{-8}$). Furthermore, we calculated the split half distance between participants (the average distance between the mean predictions of evenly splitting participants into two random groups) and found that they are highly consistent with each other (mean distance=0.50, 95% CI=[0.38, 0.71], Fig. 3.3B).

3.3 The Intuitive Physics Engine

The two experiments described previously demonstrate that across a wide range of physical scenarios, people make good predictions but are also biased in systematic ways. We argue that these predictions can be characterized using a noisy physics engine that runs probabilistic simulations – that the characteristic patterns of errors and biases we observe in participants' data can be mostly explained by uncertainty about the state of the scene after watching the video plus a noisy, approximately correct simulator that transforms those initial states into a distribution over outcomes. We formalize this hypothesis in an Intuitive Physics Engine (IPE) model.

3.3.1 The architecture of the IPE

To model people's prediction in naturalistic 3D environments, we used Unity3D as the underlying physics engine and customized to add noise to model sources of uncertainty in humans. Following K. A. Smith and Vul (2013), we considered uncertainty along three different axes (see Fig. 3.2):

Perceptual uncertainty: We modeled people's uncertainty in visual perception by adding noise to the initial positions and rotations of the objects. The starting position of each of the objects was perturbed around the true position by two-dimensional Gaussian noise parameterized by standard deviation σ_x, σ_y; and the rotation by von Mises noise around the z axis parameterized by concentration κ_z.[1]

Physical property uncertainty: We capture people's uncertainty about physical properties that vary across objects but are not directly observable – the mass of different objects – by adding Gaussian noise to the true mass, parameterized by standard deviation σ_m, truncated at zero.

Dynamic uncertainty: We considered people's uncertainty about how collision will resolve, by perturbing the resultant collision impulse force's magnitude by Gaussian noise around its true value with the standard deviation σ_F, and direction by a spherical von Mises distribution centered on the true angle of the impulse with a concentration parameter κ_θ.

[1] Following Battaglia et al. (2013), we only consider position and rotation uncertainty along a plane because most objects are resting on the ground or another object, and so uncertainty along the z-axis would cause objects to either float or interpenetrate.

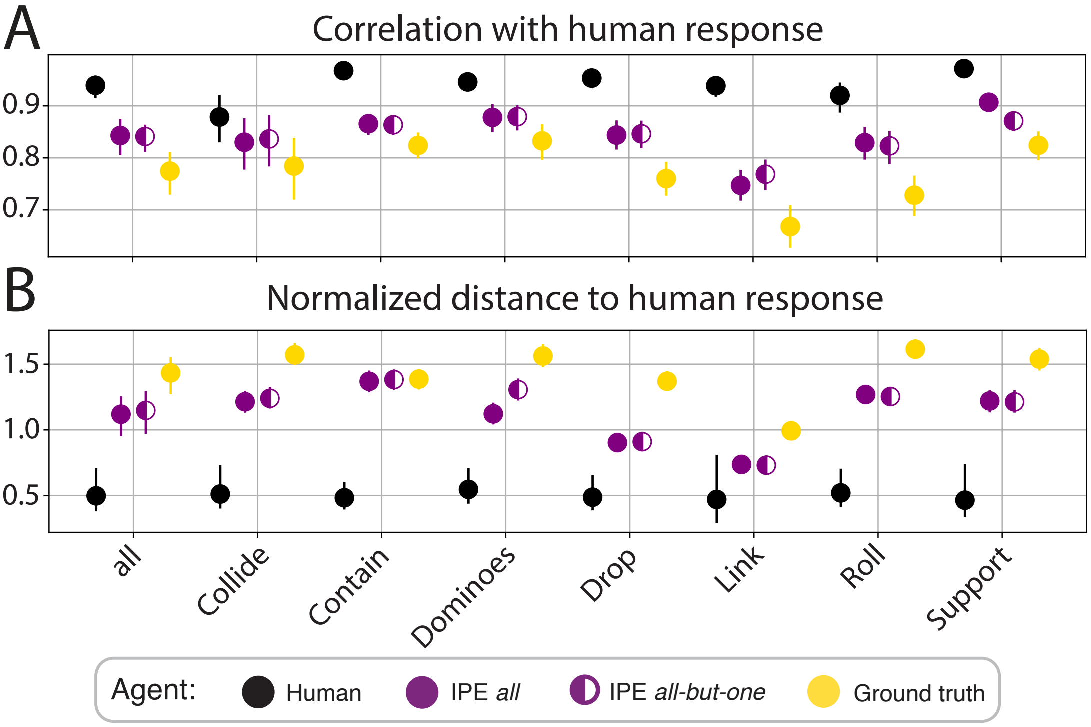

Figure 3.3. Comparison between the IPE and ground truth to human response on individual scenarios: Exp. 1 (A), Exp. 2 (B). Across all seven scenarios, the IPE better captures human response patterns than ground truth, and loses almost no predictive power across scenarios in the "all-but" fitting regime. Error bars are 95% CIs.

3.3.2 Fitting model parameters

In order to determine the set of noise parameters that best describes human behavior, we fit the six parameters defined above to participants' data on the "will it" task. Because the "where" task requires no additional modeling assumptions, we can use the same model to compare to participants' data on this task, and thus can treat model performance on that task as generalization to a separate task.

We fit parameters by simulating the scenes for 2.5s with the noisy IPE 20 times for each scene. We measured the RMSE between the proportion of IPE runs that predict contact for each trial, and the proportion of participants that do. We minimized this RMSE using the HyperOpt package (Bergstra et al., 2013).

We used two regimes for fitting. In the "all scenarios" regime, we fit the IPE to

20% of trials from each scenario (210 trials total). We then assessed performance on the 80% of trials the model had not been fit on (840 trials), testing generalization to new trials. In the "all-but-one" regime, we fit seven separate IPE models: each fit on the trials from six of the seven scenarios, then assessed performance of each of those models on the unseen scenario. Overall model performance was calculated by averaging over the performance of each model on its held-out scenario. This regime tests even stronger generalization: whether the uncertainty measured in separate scenarios can explain human predictions in scenarios uninvolved in the fitting (Wang et al., 2022).

3.4 Model Results

We first test whether a single noisy simulator can explain a range of human judgments by evaluating the IPE's predictions against humans' on both the "will it" and the "where" task. We then assess how well a set of state-of-the-art deep learning networks explain human predictions.

3.4.1 The IPE is physical-domain-general

Will it contact?

We find that a single parameterization of the IPE can explain human judgments across scenarios. Using the "all" fitting regime, the IPE achieves human-level performance across all seven scenarios on the test set (mean accuracy=0.83, 95% CI=[0.79, 0.86], Fig. 3.4A), and more importantly also has high correlation with human responses (mean correlation=0.87, 95% CI=[0.83, 0.90], Fig. 3.3A), only slightly worse than could be expected by the human noise ceiling. We also compare against how well participants would be fit by assuming perfectly accurate predictions, and find that the IPE correlates with human predictions better ($t(13)$=2.42, $p = 0.01$, Fig. 3.3A). This indicates that the IPE not only captures overall human performance, but also makes similar predictions on individual trials. Importantly, the IPE explains predictions better than the ground truth

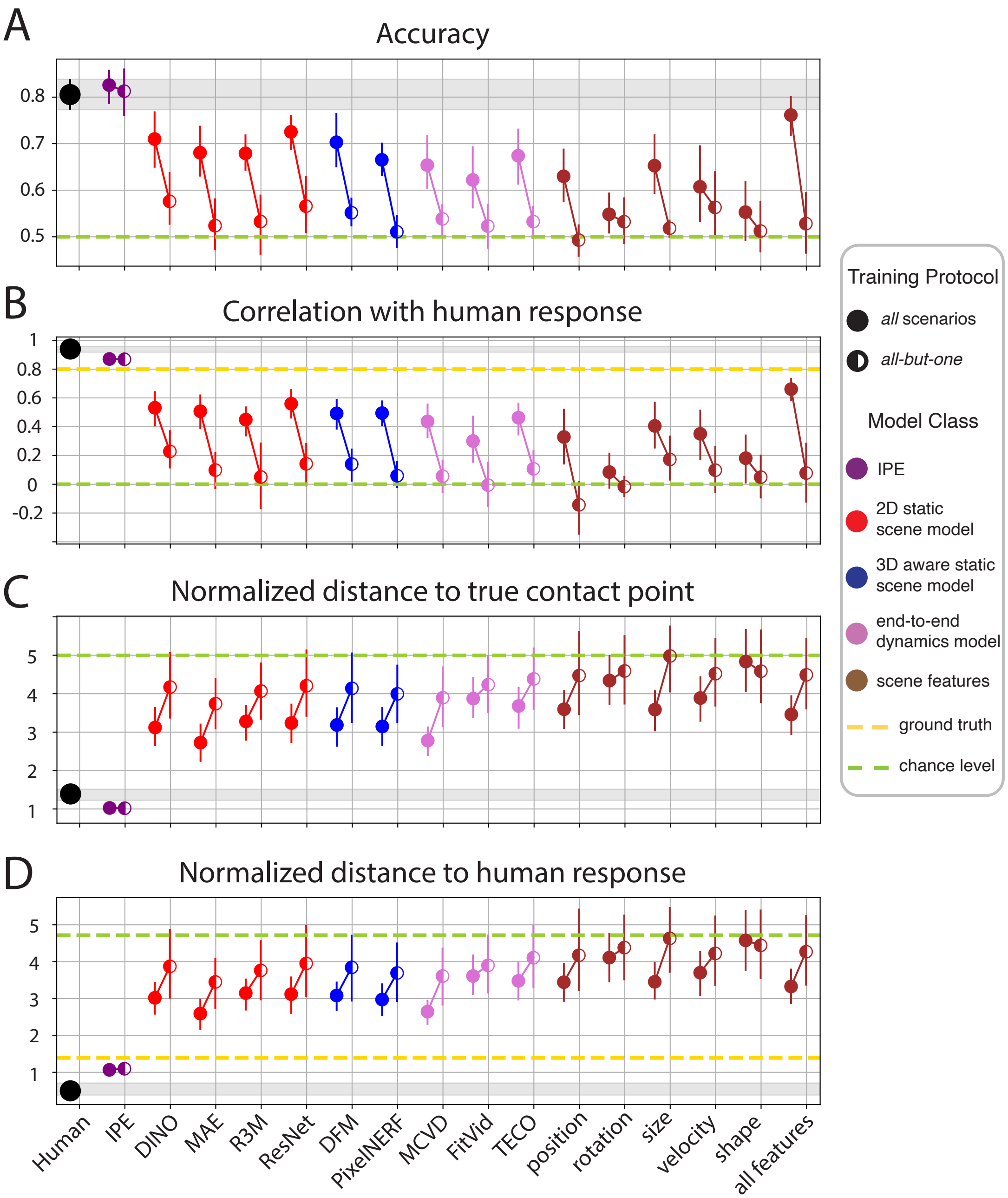

Figure 3.4. Performance of people vs. models on the two experiments. Exp 1: A. proportion of correct contact judgments, B. correlation to human responses. Exp 2: C. normalized distance to the true contact point, D. normalized distance to average human prediction. The IPE performs near human-levels across all task metrics, maintained when generalizing across scenarios. All deep learning and feature-based models perform worse than the IPE, and their performance suffers when generalizing in the "all-but-one" training protocol.

answers, suggesting that the uncertainty inherent in the model leads to uncertainty in outcomes that produce human-like errors across scenarios.

As a stronger test of generalization, we assess how well the IPE explains human data in the "all-but-one" fitting regime, where the model is assessed on scenarios it has not observed during parameter fitting. The IPE maintained high accuracy (mean accuracy=0.81, 95% CI=[0.76, 0.86]) and correlation with human responses (mean correlation=0.87, 95% CI=[0.84, 0.89]), and was nearly the same when it had access to trials from all scenarios – a pattern that held across all scenarios (Fig. 3.3A). Thus uncertainty about physical properties can be assessed in one set of scenarios and extrapolated to separate scenarios without noticeably affecting performance.

Where will it contact?

To evaluate the IPE against humans on precise location predictions, we reused the noise parameters fit on the "will it" task and extracted the contact location information between the target object and zone. The pattern of results is similar to those for the "will it" task: the IPE's distance to the true contact remained the same across fitting regimes (all: 1.03, 95% CI=[0.88, 1.21], all-but-one: 1.02, 95% CI=[0.88, 1.18]; Fig. 3.4C). The average normalized distance between the model's predictions and human placements is 1.07, 95% CI=[0.89, 1.21], significantly less than the distance between people and ground truth ($t(13)$=2.64, p = 0.01) and at the same time did not change when generalizing across scenarios (normalized distance=1.10, 95% CI=[0.91, 1.25], Fig. 3.3B). Note that the "all-but-one" fitting regime is an extremely strong test of generalization: the model must generalize across participants, physical scenarios, and even the type of prediction.

However, the IPE does capture human performance slightly worse on the "where" task than the "will it" task. This could be due to the aforementioned strong generalization hindering performance, because, e.g., one set of participants have different amounts of uncertainty than the other, or because simply judging whether contact will occur is a coarser measure than judging where contact would occur, and so parameter estimates should be less precise.

3.4.2 Comparing to deep learning and feature-based models

While we have shown that the IPE can explain human predictions across scenarios, another theory suggests human-like physics understanding can arise from less structured learning. Thus, in this section, we aim to test the generalizability of state-of-the-art deep learning models as well as models that make physical predictions based on scene features, and compare their predictions on the same stimuli to humans and IPE.

We selected state-of-the-art models from three representative model architecture classes. These models were either pretrained or finetuned on the Physion dataset. **(1)** We assessed a set of 2D static scene understanding models with an LSTM trained on Physion scenes to predict next-frame dynamics from scene representations with various amounts of structure (*DINO*, (Oquab et al., 2023); *MAE*, (He et al., 2021); *R3M*, (Nair et al., 2022); and *ResNet* (He et al., 2016)). This asseses whether the scene representations learned by these models support efficient learning of dynamics. **(2)** We assessed 3D scene understanding models with the same LSTM training on their scene representations (*DFM*, (Tewari et al., 2023), and *PixelNERF*, (Yu et al., 2021)). This assesses whether richer, 3D-aware scene representations might support better prediction. **(3)** We assess end-to-end dynamics models pretrained on Physion scenes (*MCVD*, (Voleti et al., 2022); *FitVid*, (Babaeizadeh et al., 2020); and *TECO*, (Yan et al., 2022)). These models asses whether human-like physics knowledge could be learned in an unstructured manner from video.

In order to compare deep learning models to humans on the "will it" task, for each model, we first extracted features by showing the human stimulus (450ms) and concatenated them with the "simulated" features output by the model's dynamics predictor. To get a binary output from the models, we then froze the parameters of the model and fit a logistic regression on the features. The parameters for the logistic regression were fit on a separate set of stimulus provided by the Physion dataset, with the ground truth object contact labels acting as supervision. We evaluated these models on the same unseen 840 experimental trials that the IPE was evaluated on. As seen in Fig. 3.4 AB, none of the deep learning models reached human levels of accuracy, and they did not correlate with human predictions as well as the IPE.

Similar to the IPE evaluation, we also tested the deep learning models in the "all-but-one" regime where we trained the deep models on six out of the seven scenarios and tested them on the held-out scenario. The deep models' performance dropped noticeably across the two tasks ($p < 10^{-3}$ for all models for both accuracy and correlation), with many of the models only marginally exceeding chance levels.

Next, we evaluated the deep models' predictions of where it believed contact would occur. Unlike the IPE, all the deep models compute on 2D images rather than 3D world coordinates, so we used logistic regression on the same model features as before to output a prediction probability distribution on a 16×16 grid over the image and then projected the center of each cell in the grid to 3D world coordinates (see Fig. 3.5).[2]

In order to compare between humans, the IPE, and deep learning models, we needed to align their predictions. First, we transformed human and IPE predictions by translating their predictions on world coordinates back into 2D image coordinates, similarly binning them into 16×16 grids, approximating the prediction using center of the cell and then projecting the center back to 3D world coordinates. Second, because participants and the IPE were only allowed to make predictions on the zone area, we re-normalized the prediction probability distribution from the deep learning models to be only on the zone. We then measured probability-weighted distance between the grid center points on world coordinates as a metric for prediction distance.[3] As shown in Fig. 3.4 CD, the

[2] We found empirically that structuring our problem as a 16×16 grid classification task improved the readout training performance compared to a position regression task, while also guaranteeing fine-grained predictions.

[3] We also considered Wasserstein distance over grid distributions, and found qualitatively similar patterns of results.

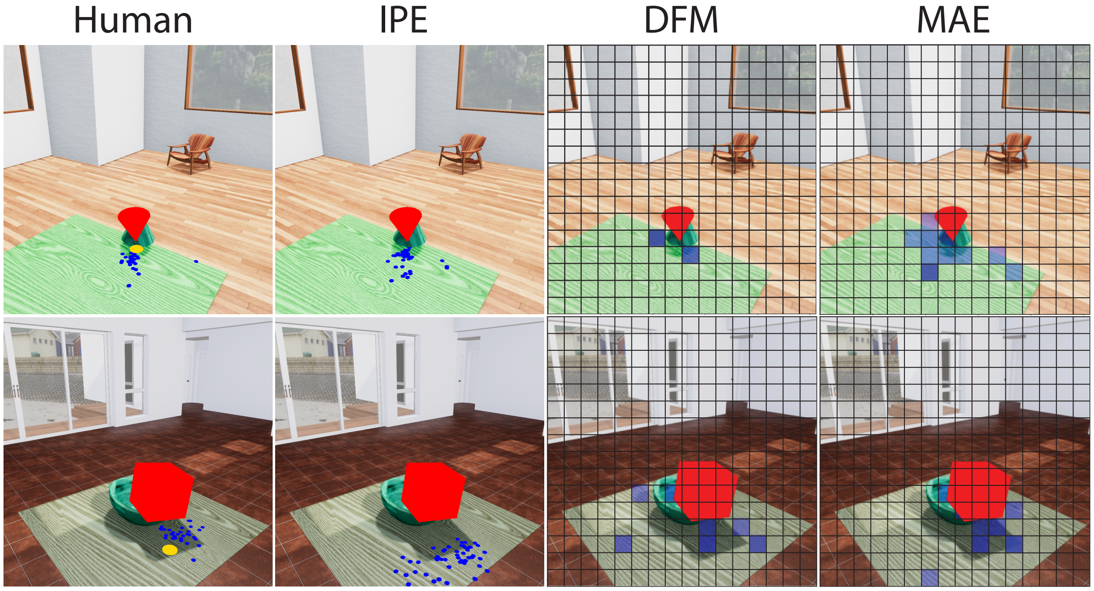

Figure 3.5. Two example "where" trials showing people's and models' predictions. The target object is highlighted in red, predictions are colored in blue. The ground truth contact point is colored in gold in the human panel.

deep learning models did not capture human performance as well as the IPE, and always had decreased predictivity in the "all-but-one" regime, though here most performed above chance, providing evidence that they had learned something about the physics of these scenes as a whole.

To evaluate feature-based models on the same tasks, we used objects' position, rotation, size, shape and velocity at the stopping frame (i.e. 450ms) as features and fit a linear readout model in the same way we did for deep learning models for each feature. Additionally, we also tested the combination of all the features. We found that featured-based models performed much worse than humans and IPE, especially when they were asked to generalize to unseen scenarios (Figure 3.4).

3.5 Discussion

In this paper, we tested the hypothesis that human physical predictions can be explained by approximate probabilistic inference in a single, general physics simulator across a wide range of everyday settings. Across two experiments, we found that a physics engine that runs probabilistic simulations generalized to unseen stimuli in human-like ways, but a set of state-of-the-art deep learning models and feature-based models do not yet reach that level.

One major point of difference between the IPE and the deep learning models we tested here is the input encoding: the IPE takes 3D information of the scene as inputs whereas the deep learning models compute on pixels. Learning directly from pixels can allow for greater flexibility in the representation of scenes and dynamics, but imposes the challenge of learning to extract scene information rather than that information being provided. In this case, however, it appears these representations do not support longer term predictions in untrained scenarios that require understanding the physics of the world. In future work, we will evaluate a broader set of deep learning models that impose greater structure on the learning of physics – e.g., graph neural networks that work from scene representations and explicitly parse the world into objects and their relations (Allen et al., 2022; Battaglia et al., 2016; Han et al., 2022; Li et al., 2018; Mrowca et al., 2018) – as well as noisy simulation models that rely on scene parsing models to provide information about the world (Wu et al., 2017). Systematic testing of broader sets of models can help inform us what additional structure is required to develop more human-like understandings of the physical world.

While the IPE model predicts human response patterns well, it is still below the noise ceiling. This is a pattern found in many studies going back to Battaglia et al. (2013), and is likely due to the fact that humans cognitive simulations are not exactly the same as computer physics engines, but instead have different implementations and additional simplifications (Bass et al., 2021; Chen et al., 2023; Li et al., 2023). Further research into the structure of human physical representations and simulations will be required to close this gap.

The physical world is complex and open-ended, yet we easily reason about a wide range of scenarios that we might encounter in everyday life. The current study suggests that this robust generalization behavior often comes from having a generalizable mental model of the physical world and the ability to continuously simulate forward about how the world will unfold. In the long term, such studies will help us to understand and implement the computational mechanisms needed for the deep learning models to be more human-like.

3.6 Acknowledgments

Chapter 3, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Wang, H., Jedoui, K., Venkatesh, R., Binder, F., Tenenbaum, J., Fan, J., Yamins, D., & Smith, K. (2024). Probabilistic simulation supports generalizable intuitive physics. In *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*. The dissertation author was the primary investigator and author of this material.

References

Allen, K. R., Rubanova, Y., Lopez-Guevara, T., Whitney, W., Sanchez-Gonzalez, A., Battaglia, P. W., & Pfaff, T. (2022). Learning rigid dynamics with face interaction graph networks.

Babaeizadeh, M., Saffar, M. T., Nair, S., Levine, S., Finn, C., & Erhan, D. (2020). Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413–424.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLOS Computational Biology*, *15*(7), e1007210. https://doi.org/10.1371/journal.pcbi.1007210

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.

Battaglia, P. W., Pascanu, R., Lai, M., Jimenez Rezende, D., & Kavukcuoglu, K. (2016). Interaction Networks for Learning about Objects, Relations and Physics. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 4502–4510). Curran Associates, Inc.

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., et al. (2021). Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *International conference on machine learning*, 115–123.

Chen, T., Allen, K. R., Cheyette, S. J., Tenenbaum, J., & Smith, K. A. (2023). " just in time" representations for mental simulation in intuitive physics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*. https://doi.org/10.1037/rev0000281

Gilden, D. L., & Proffitt, D. R. (1989). Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 372.

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.

Han, J., Huang, W., Ma, H., Li, J., Tenenbaum, J., & Gan, C. (2022). Learning physical dynamics with subequivariant graph neural networks. *Advances in Neural Information Processing Systems*, *35*, 26256–26268.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked autoencoders are scalable vision learners. *arXiv:2111.06377*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kubricht, J. R., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., & Lu, H. (2016). Probabilistic Simulation Predicts Human Performance on Viscous Fluid-Pouring Problem. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*.

Li, Y., Wang, Y., Boger, T., Smith, K. A., Gershman, S. J., & Ullman, T. D. (2023). An approximate representation of objects underlies physical reasoning. *Journal of Experimental Psychology: General*.

Li, Y., Wu, J., Tedrake, R., Tenenbaum, J. B., & Torralba, A. (2018). Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*.

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fei-Fei, L. F., Tenenbaum, J., & Yamins, D. L. (2018). Flexible neural representation for physics prediction. *Advances in neural information processing systems*, 8799–8810.

Nair, S., Rajeswaran, A., Kumar, V., Finn, C., & Gupta, A. (2022). R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*.

Nayebi, A., Rajalingham, R., Jazayeri, M., & Yang, G. R. (2023). Neural Foundations of Mental Simulation: Future Prediction of Latent Representations on Dynamic Scenes.

Neupärtl, N., Tatai, F., & Rothkopf, C. A. (2021). Naturalistic embodied interactions elicit intuitive physical behaviour in accordance with Newtonian physics. *Cognitive Neuropsychology*, 1–15. https://doi.org/10.1080/02643294.2021.2008890

Nusseck, M., Lagarde, J., Bardy, B., Fleming, R., & Bülthoff, H. H. (2007). Perception and prediction of simple object interactions. *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization*, 27–34.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., . . . Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision.

Piloto, L. S., Weinstein, A., Battaglia, P. W., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 1–11. https://doi.org/10.1038/s41562-022-01394-8

Proffitt, D. R., Kaiser, M. K., & Whelan, S. M. (1990). Understanding wheel dynamics. *Cognitive psychology*, *22*(3), 342–373.

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

Smith, K., Hamrick, J., Sanborn, A. N., Battaglia, P., Gerstenberg, T., Ullman, T., & Tenenbaum, J. (2024). Intuitive physics as probabilistic inference.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Tewari, A., Yin, T., Cazenavette, G., Rezchikov, S., Tenenbaum, J. B., Durand, F., Freeman, W. T., & Sitzmann, V. (2023). Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *NeurIPS*.

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.

Voleti, V., Jolicoeur-Martineau, A., & Pal, C. (2022). Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. *(NeurIPS) Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2205.09853

Wang, H., Allen, K. R., Vul, E., & Fan, J. E. (2022). Generalizing physical prediction by composing forces and objects. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Wu, J., Lu, E., Kohli, P., Freeman, W. T., & Tenenbaum, J. B. (2017). Learning to See Physics via Visual De-animation. *Neural Information Processing Systems*, 12.

Yan, W., Hafner, D., James, S., & Abbeel, P. (2022). Temporally consistent transformers for video generation. *arXiv preprint arXiv:2210.02396*.

Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021). pixelNeRF: Neural radiance fields from one or few images. *CVPR*.

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023). Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*.

Chapter 4

Discussion

What are the underlying representations and computations that enable us to robustly generalize? In this dissertation, I propose that prior theories of human generalization differ in their assumptions about how abstractly past experiences is represented, and that different theories lie on different levels of a "ladder of abstraction". At the bottom level of the ladder is exemplar-based approaches, where concrete experience is stored into memory and when generalizing, it is compared against the novel situation. Heuristics is one level up on the ladder, assuming summary statistics or particular features extracted from past experiences should suffice when generalizing to new environments. On the top level of the ladder is the world model hypothesis, which posits that people have an internal causal model of the external world, incorporating knowledge about objects, events, and the relationships between them. A review of the literature suggests that different generalization behaviors across and within domains are explained by different theories, we lack a coherent account of how people can make accurate judgements and predictions in novel environments.

Across three chapters, I systematically examine how people generalize to new situations by designing various scenarios and manipulating the stimuli distribution. I show that the world model hypothesis is a promising candidate of being a unifying account of people's generalization behaviors across and within domains. Specifically, in Chapter 1, I show that in the domain of visual concept learning, both people's memorization of specific units and learning abstract rules can be explained by a program learning model that exploits the statistical regularities in the stimuli, which also explains how people can flexibly adapt to new objects not experienced before. Chapter 2 focuses on how people understand the relationships between different concepts and make physical predictions in a novel situation. I show that people achieve high generalization performance, in part, by constructing composable internal models of the physical scene and performing model-based compositional generalization. Chapter 3 builds on the results from Chapter 2 by asking people to generalize to not just one, but many novel physical contexts. I show that people's robust generalization across multiple scenarios is most consistent with the predictions from an intuitive physics engine that can represent different objects and properties, but runs probabilistic simulations on finite computational resources. Altogether, this dissertation brings us closer to understanding the underpinnings of humans' robust, flexible behavior when faced with novel situations. My research suggests that representing entities and how these entities interact in an abstract world model plays a key role and potentially provides a more coherent account of human generalization.

While this dissertation thus far has focused on how these abstract world models facilitates generalization for adult participants, one question that arises from this work is where such structured representations come from in the first place. There has been a wealth of research from the developmental literature on how infants acquire new physical concepts as they develop (Baillargeon, 1994; Baillargeon et al., 1995; Kuhn, 2012), however, it is less clear what underlying cognitive mechanisms are driving this continual learning behavior. One important future direction is to reverse-engineer the underlying learning algorithms by building a computational model that can learn from the same data infants are exposed to and explain their learning trajectory (Vong et al., 2024).

From another perspective, the learning of world models does not only take place in infancy: our mental model of the world is not static, rather, it is dynamic and continuously updated through experience. One key factor that is guiding the updating of world model is curiosity (Chu & Schulz, 2020; Martinez et al., 2023). Curiosity plays a key role in exploration and the acquisition of new information. Future investigations could examine what is the specific curiosity signal that's driving exploration in the physical world, and in what way curiosity interacts with the refinement of internal representations.

Another way that world models infiltrate our everyday life is through communication. When we talk, we are attempting to communicate ideas about the objects and events that take place around us. Another key direction for future work is to move these findings beyond the individual level towards interaction between individuals and even larger communities. Recent work has begun to study how such representations of the world can be shared between people through communication to enable improved collaboration (McCarthy et al., 2021), trust (Brockbank et al., 2022), and cultural transmission of ideas (Wang et al., 2022). Understanding how these fit together informs not only cognitive science and psychology, but also provides guidance for robotics and artificial intelligence.

References

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, *3*(5), 133–140.

Baillargeon, R., Kotovsky, L., & Needham, A. (1995). The acquisition of physical knowledge in infancy.

Brockbank, E., Wang, H., Yang, J., Mirchandani, S., Bıyık, E., Sadigh, D., & Fan, J. E. (2022). How do people incorporate advice from artificial agents when making physical judgments? *arXiv preprint arXiv:2205.11613*.

Chu, J., & Schulz, L. E. (2020). Play, curiosity, and cognition. *Annual Review of Developmental Psychology*, *2*, 317–343.

Kuhn, D. (2012). The development of causal reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 327–335.

Martinez, J., Binder, F., Wang, H., Haber, N., Fan, J., & Yamins, D. L. (2023). Measuring and modeling physical intrinsic motivation. *arXiv preprint arXiv:2305.13452*.

McCarthy, W. P., Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *arXiv preprint arXiv:2107.00077*.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511.

Wang, H., Yang, J., Tamari, R., & Fan, J. E. (2022). Communicating understanding of physical dynamics in natural language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).