From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Neuroscience

2021

Learning intuitive physics: a computational investigation

Piloto, Luis Rolando

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp013t945t905

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Learning Intuitive Physics: A Computational Investigation

Luis Rolando Piloto

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Princeton Neuroscience Institute
Adviser: Matthew M. Botvinick

June 2021

© Copyright by Luis Rolando Piloto, 2021.

All rights reserved.

Abstract

Intuitive physics is the ability to reason about three-dimensional objects, their dynamic interactions, and responses to forces. The development of intuitive physical knowledge in humans has been extensively catalogued by developmental psychologists in order to study the "origins of knowledge." It is at the center of a debate, parallel to that of language, between nativists and empiricists. Nativists have demonstrated evidence of physical intuition as early as 2.5 months of age (by some accounts even younger) suggesting domain-specific constraints on learning. Empiricists have shown (for some domains) that general-purpose learning principles, instantiated via artificial neural networks, are capable of producing the stage-like progression of rule-like behavior seen in infants. However, methodological limitations abound on both sides, impeding progress on this dialog. Nativists have experimental limits on probing newborns and young infants due to perceptual and attentional constraints. Empiricists have only been able to build learning systems for gross simplifications of the physical world, abstracting sensory information to physical quantities and learning based on that.

The main contribution of this thesis is to capitalize on advances in deep learning to build models that can *actually* learn intuitive physics from visual data. In doing so, this work unlocks a necessary component for mapping the mechanisms of intuitive physical knowledge: *any* model that seeks to provide a mechanistic account of learning phenomenon must first be able to learn that phenomenon. We adopt the Violation-of-Expectation paradigm used in developmental psychology to probe infant physical intuition and build a synthetic dataset to probe physical knowledge in deep neural networks. In Chapter 2, we build a model with a domain-general architecture and find it only capable of narrow generalization of intuitive physics knowledge. In Chapter 3, we build a model with an object-centric inductive bias (reflecting hypothesized domain-specific constraints) and find it is capable of learning generalized

physical concepts with as little as 28 hours of visual experience in our synthetic 3D environment. Furthermore, we ablate the architecture and find the object-centric components are necessary for the rich generalization. Finally, Chapter 4 briefly elaborates potential mechanisms for learning necessary components of the object-centric model: object discovery and tracking.

Acknowledgements

Countless people helped me complete this thesis. I will now attempt to count them. First and foremost is my advisor Matt Botvinick. He was generous with his time and attitude. And flexible with how he supported me: toeing the line between encouragement and cajoling as needed. I benefited from his wisdom, advice, and honesty for years (and if I have anything to say about it, I'll continue to do so). Many times I would present Matt with an either-or dilemma and he had this superpower to magically pull a win-win rabbit out of the hat. And throughout the difficult progression of a thesis project that at times felt insurmountable, I always felt like Matt had my back. Thank you, Matt.

Ken Norman was my advisor before I moved to DeepMind. And before that, Ken hired me as a research assistant. His impact on my academic identity was foundational. Throughout that time (and even beyond), Ken showed me unwaivering support in such a genuine way that never felt mechanical or stemming merely from a sense of duty as my supervisor. Besides scientific knowledge, I also absorbed Ken's views on *how* to do science: openly and with rigour and compassion. Thank you, Ken.

I am grateful to the Casey Lew-Williams for helpful feedback on thesis. I am grateful to Alexandra Michaud for helping in so many ways get through the logistics of completing a Ph.D. I am grateful to the PNI faculty for creating such a thoroughly collaborative and wonderful environment for scientific research.

At DeepMind, I am grateful for the various collaborators and colleagues who made this thesis possible either through direct or indirect contributions. Ari Weinstein and Dhruva Tirumula were absolutely critical in developing the first versions of our intuitive physics dataset. Both Ari and Dhruva were so delightful to work with and casually hilarious each in their own way. It made those early days of the project much easier. I'm additionally grateful to Ari who has provided a wealth of personal advice

throughout my time at DeepMind. He always leverages perspective to comedic effect and actionable insight. Chris Burgess, Loic Matthey, Rishabh Kabra, and Alexander Lerchner were a steady source of inspiration and feedback. They helped me form intuitions about information theory, variational auto-encoders, generative modelling, and objects that form a major component of my current scientific perspective. In Toni Creswell, I found an academic ally to proselytize the alignment problem and together we laid the foundations of the AlignNet work. When working with Peter Battaglia I've always been impressed by his full-stack expertise: he's frequently insightful at high-level research questions as well as low-level technical details. I am grateful for his guidance on my thesis project and as my manager and for granting me the time to finish writing this dissertation. Finally, I am grateful to DeepMind for providing such a wonderful research environment.

I am grateful for the friendship and support of my PNI cohort. I benefited immensely, both scientifically and personally, from being a member of the Computational Memory lab. Working with James Antony was a delight and came with the added bonus of his friendship. I am grateful for Jamal Williams's constant comradery. Alex Piet brought an infectious silliness to our apartment that got me through many stressful days in lab. Ben Deverett was my roommate at Princeton and a long-time friend and I am so grateful to have gotten to spend that much time in the same state as him. I think I may still owe him $10. Leaving my friends and family in NJ to move to London was difficult. I was sustained by early friendships that have continued to this day: Jacob Byrne, Bryan Adare, Jon Voutila, Sam Lai, and Kerry Klemmer.

Finally, I am grateful for the support of my family throughout my life. My brother, Obdulio Piloto, served as a role model in science and in life and encouraged me from an early age. My sisters Liz, Lili, and Nat made many sacrifices so that I could be in the position that I am today. Liz taught me to be brave, Lili taught me to never give up, and Nat taught me to be kind. My mother was a constant cheerleader and

served as a model of perseverance. In London I met my wife, Livia. There's no way I can put into words all the ways Livia has helped and inspired me, but I am prepared to quantify it: She has been a part of less then 14% of my life, yet is a (figurative) co-author on 100% of my proudest achievements. My daughter Laila has energized me to after long days to wake up and do it all again. Even my unborn daughter, Lana, has helped: her upcoming birth served as the catalyst for me to buckle down and write this thesis.

To my wife, Livia Piloto.

Contents

Abstract iii

Acknowledgements v

List of Tables xiii

List of Figures xiv

1 Introduction **1**

1.1 Motivation 1

1.2 Overview of Thesis 5

1.3 Intuitive Physics In Infants 6

1.3.1 Early Probes of the Infant's Knowledge of Intuitive Physics . 6

1.3.2 The Violation-of-Expectation Paradigm 7

1.4 Intuitive Physics And Artificial Systems 13

1.4.1 Learning Physics 13

1.4.2 Discovering Objects 14

2 Learning Intuitive Physics Without Objects **18**

2.1 Abstract 20

2.2 Introduction 20

2.3 The Violation of Expectation Paradigm 23

2.4 Dataset and Probe Design 23

2.4.1 General Dataset Design 24

2.4.2 'Object Persistence' . . . 26
2.4.3 'Unchangeableness' . . . 27
2.4.4 'Continuity' . . . 27
2.4.5 'Solidity' . . . 28
2.4.6 'Containment' . . . 29
2.5 Model . . . 29
2.5.1 Training and Evaluation . . . 31
2.6 Results . . . 32
2.7 Discussion . . . 34
2.7.1 Present Work . . . 34
2.7.2 Future Work . . . 35
2.8 Appenfix . . . 41
2.8.1 Convolutional Architecture . . . 41
2.8.2 Model Representations For Consistent Vs. Inconsistent Probes 41
2.8.3 Model Predictions for Inconsistent Probes . . . 43

3 Learning Intuitive Physics With Objects 46

3.1 Introduction . . . 48
3.2 Results . . . 58
3.2.1 Models . . . 58
3.2.2 Training and Testing . . . 60
3.2.3 Test set performance . . . 62
3.2.4 Effect of training set size . . . 64
3.3 Discussion . . . 65
3.4 Materials and Methods . . . 70
3.4.1 Training Dataset: 'Freeform' Physical Events . . . 70
3.4.2 Violation Of Expectation (VoE) Paradigm . . . 71
3.4.3 Physical Concept: 'Object Persistence' . . . 73

3.4.4 Physical Concept: 'Unchangeableness' . . . 73
3.4.5 Physical Concept: 'Continuity' . . . 73
3.4.6 Physical Concept: 'Solidity' . . . 74
3.4.7 Physical Concept: 'Directional Inertia' . . . 75
3.4.8 Model Architecture . . . 75
3.4.9 Perception Module . . . 76
3.4.10 Dynamics Predictor . . . 77
3.4.11 Baseline Models . . . 79
3.4.12 Experiments . . . 80
3.4.13 Visual Experience Calculation . . . 80

4 Object Representation and Tracking 82

4.1 Slow Multi-object Network . . . 83
4.1.1 Introduction . . . 83
4.1.2 Methods . . . 85
4.1.3 Results . . . 91
4.2 AlignNet . . . 92
4.2.1 Introduction . . . 92
4.2.2 Methods . . . 96
4.2.3 AlignNet V1: Learning Similarity . . . 98
4.2.4 AlignNet V1.5 . . . 104
4.2.5 Conclusion . . . 106

5 Conclusion 108

5.1 Summary of Thesis Work . . . 108
5.2 Future Work . . . 110
5.2.1 End-to-end, Unsupervised Learning of Intuitive Physics . . . 110
5.2.2 Expanded Physical Concepts . . . 111

5.2.3 Real-world Data For Deep Neural Networks 111
5.2.4 Synthetic Data For Real-world Babies 112
5.2.5 Manipulating and Measuring the Statistics of Visual Experience 113

Bibliography **115**

List of Tables

List of Figures

2.1 Example (subsampled) test probes for each dataset. The first and second rows show consistent and inconsistent probes, respectively. Categories: *a*, Object Persistence; *b*, Unchangeableness; *c*, Continuity; *d*, Solidity; *e*, Containment 36

2.2 Examples of (subsampled) training data (consistent examples and/or controls). Categories: *a*, Object Persistence; *b*, Unchangeableness; *c*, Continuity; *d*, Solidity; *e*, Containment 37

2.3 Model architecture for a single time step. Dotted lines depict sampling from a distribution and dashed lines denote a one time-step delay. Thus, the sampled z_t and the memory state from the *previous* time step are used to calculate the prior and posterior for the current time step. 38

2.4 Difference in KL divergence (x-axis) over final evaluation set of 8,000 probe pairs. Histograms show the distribution over the evaluation set with outliers >3 standard deviations removed. Text shows the result of significance tests with an asterisk denoting significance. Red corresponds to models trained separately, blue to model trained concurrently. 38

2.5 Reliability of Relative Surprise (y-axis) for models trained with different Training Set Sizes (x-axis). Each point represents a different evaluation set of 2,000 probe pairs with green points showing a statistically significant effect. As the training set size increases, the variance of the effect size decreases and reliability of statistical significance increases. 39

2.6 Top row: Subsampled frames from an inconsistent probe for the Object Persistence dataset. Here physics is violated when the sheet magically collapses the magenta cube. Bottom row: Network's predictions for the above frame based only on previous frames. Note how it rectifies the input with the correct physics displaying a scene where the sheet rests on top of the magenta cube. 39

2.7 Relative Surprise Over Training for Models Trained on Datasets Individually (dashed lines) or Simultaneously (solid lines) 40

2.8 t-SNE Embeddings 42

2.9 Network predictions (bottom row) for Continuity inconsistent probe (top row). Even though the inconsistent probe has the ball disappear between the pillars, the network predicts it should appear. 43

2.10) Network predictions (bottom row) for Unchangeableness Position inconsistent probe (top row). Even though the object positions have been swapped after the occluder is lifted, the network predicts the objects should keep their initial placement. 43

2.11) Network predictions (bottom row) for Unchangeableness Color inconsistent probe (top row). Even though the cube has changed color after the occluder is lifted, the network predicts it should keep the original color. 44

2.12) Network predictions (bottom row) for Unchangeableness Shape inconsistent probe (top row). Even though the yellow sphere has been changed to a yellow cube after the occluder is lifted, the network predicts it should remain a sphere. 44
2.13) Network predictions (bottom row) for Object Solidity inconsistent probe (top row). Even though the cube falls through the container and floor in the inconsistent probe, the model correctly predicts it should remain in the container. 44
2.14) Network predictions (bottom row) for Containment inconsistent probe (top row). When the container is moved, the cube that fell into it magically phases through it, remaining in the place the container used to be. In the first frame after this manipulation, the model correctly predicts nothing should be in the place where the container used to be. However, the model cannot hold onto this correction for an extended period of time and integrates the inconsistent information into the prior, as seen in the final frame where the space that should be empty has a blurred entity in it. 45

3.1 Example probes adapted from (Aguiar and Baillargeon, 1999) to assess the physical concept of *continuity* (Spelke et al., 1992; Baillargeon, 2002): objects trace a continuous path through time and space. Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), a ball rolls behind two occluders. In the second possible probe (second row), no ball is present. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the ball rolls behind the first occluder and emerges from the second occluder, never appearing between the two occluders. The second impossible probe (fourth row) has the opposite structure: the ball appears between the two occluders, but was never seen rolling behind the first occluder or rolling out of the second occluder. 51

3.2 Example probes adapted from (Baillargeon, 1987) to assess the physical concept of *object persistence* (Baillargeon, 2008). Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), a rigid plank falls onto, occludes, and is propped up by a cube. In the second possible probe (second row), the plank falls and remains flush against the ground. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the plank falls onto the cube. However, instead of being propped up by the cube as you would expect, the plank ends up flush against the ground as if no cube existed at all. The second impossible probe (fourth row) shows a plank falling where there is no cube. Although it should end up flush against the ground, the plank comes to rest at an angle as if it landed on a cube. 52

3.3 Example probes adapted to assess the physical concept of *solidity* (Hespos and Baillargeon, 2001c): objects do not interpenetrate. Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), a tall block falls into a visibly empty container. In the second possible probe (second row), a short block falls into the container. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the tall block continues to fall into the container as if there was no bottom. In the second impossible probe (fourth row), the short block falls into the visibly empty container but fails to fall all the way in. 53

3.4 Example probes adapted from (Káldy and Sigala, 2004) to assess the physical concept of *unchangeableness* (Baillargeon and Carey, 2012b; Spelke et al., 1992): objects have certain immutable attributes (e.g. they do not change shape, size, or color). Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), an arrangement of objects is occluded by a descending curtain and eventually unoccluded to reveal the initial configuration of objects. In the second possible probe (second row), the same exact objects appear in a different configuration. They are then occluded and unoccluded as in the first probe. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the objects begin in the configuration of the first probe. However, when the occluder is lifted, the objects are seen in the configuration of the second probe, suggesting that the objects have somehow changed shape, size, or color while occluded. The second impossible probe (fourth row) starts with the configuration from the second possible probe and ends with the configuration from the first possible probe. . 55

3.5 Example probes adapted from (Spelke et al., 1994a) to assess the physical concept of *directional inertia*: colliding objects undergo changes to the direction of their velocity in accordance with the angle of incidence. Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), a ball rolls towards a block, ricochets off the block (which reverses its horizontal velocity) and maintains its movement towards the rear of the scene. The second possible probe (second row), reverses this trajectory: it starts at the rear of the scene, ricochets horizontally off the block, and continues to roll towards the front of the scene. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the ball rolls from the front of the scene towards the rear. When it collides with the block, instead of continuing towards the back of the scene, it reverses its direction back to the front of the scene. Similarly, in the second impossible probe (fourth row), when the ball collides with the block, it unexpectedly rolls back in the direction it came from. 56

3.6 Example training videos (temporally downsampled, actual videos contain 15 frames). Scenes are constructed procedurally with composable interactions: objects added to a scene are either initialized completely randomly or target preexisting objects in the scene. Camera starts from a fixed location and drifts randomly over time. 57

3.7 Model architecture consists of two components: the perception module (left) and the dynamics predictor (right). The perception module takes a segmented image $(x^t, m^t_{1:K})$ and is pre-trained to reconstruct each object individually (left) using a variational autoencoder with shared parameters: Φ, θ. In doing so, Φ can produce a set of object codes, $z^t_{1:K}$, which represent static object properties in a learned representational space. For training the dhynamics model, the object codes are aggregated over the course of a video into an object buffer. An interaction network (IN) then computes interactions from an object-based LSTM to the object buffer and the current frame's object codes. The result of these interactions are used to update the LSTM and make per-object predictions. 59

3.8 Illustration of the splicing procedure used to create a probe tuple for the physical concept of 'unchangeableness' as seen in Figure 3.4. Probe tuples consist of two possible videos and two impossible videos. The possible videos are created by running a physics engine with related initial conditions (e.g. same objects but different positions). Importantly, videos are constructed so they share a frame in common (this is the center image in the figure above). The segments before this shared frame are called the "start segment" (left part of image) and the ones after are called the "end segment" (right part of image). Thus to create any probe, we can gather frames from a start segment, add the common frame, and add frames from an end segment. Possible probes (green arrows above) are formed by starting and ending at corresponding start and end segments (e.g. start segment A to end segment A). Impossible probes (red arrows above) are formed by using a mismatched end segment (e.g start segment A to end segment B). This design ensures that the possible videos and impossible videos within a probe tuple are perfectly matched in terms of individual frames and pairs of frames. 61

3.9 **Top, Middle Rows**: Comparison of object-based model, PLATO (green) against flat baseline models (blue). PLATO consistently shows large effect sizes (top row) in response to impossible events and robustly classifies them as impossible (center row). Even when the flat model has more model parameters (dark blue), it only shows above-chance accuracy on two of the five concepts. Each point represents the average performance over 5,000 probe tuples for models trained with a different random seeds. Error bars show the 95% confidence interval over five seeds. **Bottom:** Frame-by-frame analysis of surprise reveals that the relative surprise increases substantially when videos become physically impossible. Relative surprise computed at each frame of probe videos for five random seeds of the PLATO model. The x-axis indicates frame index within a video. Vertical lines indicate when the impossible probes become aphysical (varies by dataset). Blue and red dots code points above or below zero respectively - only applied to frames after the onset of aphysicality. Dots (small) depict different random seeds. Dots (large) depict mean over three seeds with corresponding errorbars showing 95% confidence interval over seeds. . . . 63

3.10 PLATO shows robust VOE effects when the dynamics predictor trained with as little as 50,000 examples (28 hours of visual experience). Size = '0K' indicates no training at all and shows that object representations are not sufficient for VOE effects, training the dynamics predictor is critical. Performance was averaged across all five physical concepts (see Supplementary Material for results for broken down by individual concepts). Dots (small) depict different random seeds. Dots (large) depict mean over three seeds with corresponding errorbars showing 95% confidence interval over seeds. 64

4.1 SlowMONet successfully uses motion information to track identical objects in a simple multiple object tracking dataset. **Left Column:** The top row shows an image sequence of ten frames, which starts at the top-left going rightward and wraps around after five frames. The sequence consists of three identical green squares drifting within a bounded square region. Below the input sequence we show the reconstructed object for each object slot. We can see that despite having identical visual features, SlowMONet keeps each object in the same object slot throughout the sequence. **Right Column:** The top row shows another input example consisting of identical blue squares. The second object slot reconstruction shows that despite stably tracking all visible objects - the model's representation of perfectly occluded objects does not persist longer than one frame. 92

4.2 SlowMONet applied to Physical Concepts dataset. The sequence of images is shown starting at the top-left and wraps around to the next row. **Left:** The input data depicts two containers (one green, one pink) and a yellow ball. Additionally, a box, that is only ever partially in-view, falls from above the viewport, hits the corner of the green container, and bounces out of frame again. **Center:** SlowMONet's reconstruction of the image sequence. The reconstruction is fairly accurate, but a bit blurry. Some colors (i.e. specular highlight on the ball) are muted. **Right:** The segmentation masks learned by SlowMONet. Each unique color corresponds to an inferred object. The green container (assigned the light blue color slot) is correctly segmented in the same slot throughout the entire sequence. An impressive result from this example is that the falling box (assigned the red color slot) is correctly segmented into a stable slot through the sequence. Errors in segmentation are seen where the model over-segments the checkerboard floor, assigning one object slot to each of the. Errors in the stability of segmentation are seen in the segmentation of the ball which is initially assigned the purple slot but after being completely occluded in the fourth frame emerges again in a different slot (the green slot assigned to part of the checkerboard). 93

4.3 Classification accuracy for aligning with various amounts of noise (controlled by ϵ) in the Symbolic Objects dataset. We compare the baseline, no-learning approach "(MSE)" against the AlignNet. Even at the highest noise level, both models achieve high accuracy. 100

4.4 **Top:** Illustration of a two-frame sequence with a green ball and a blue ball in each frame. **Bottom**: Without assuming any structure in how objects are allowed to transform while retaining identity, the correct alignment is ambiguous. The red arrow suggests aligning to keep color constant. The purple arrow suggests aligning to keep position constant. 101

4.5 Classification accuracy on a simulated version of the structured transformation problem using the Symbolic Objects data for increasing values of ϵ. The AlignNet can learn to attend to unchanging features of an object to solve the alignment almost perfectly event at high levels of noise which degrade the baseline model's (MSE) performance. . . . 102

4.6 Classification accuracy on a simulated version of the structured transformation problem using the Learned Objects dataset for increasing values of ϵ. Note that due to multiple potential valid alignments over empty objects, the baseline accuracy without any noise is around 60. As seen in the Symbolic Objects dataset, the AlignNet can tolerate much higher levels of noise in comparison to the baseline model (MSE). 103

4.7 Classification accuracy for aligning across various temporal gaps (controlled by δ) in the Learned Objects dataset. We compare the baseline, no-learning approach "(MSE)" against two versions of the AlignNet. AlignNet (v1) is fundamentally incapably of handling the ECC problem, but still fairs better than the baseline as δ increases. Adding the second attention step in AlignNet (v1.5) substantially improves performance for the maximum value of $delta = 8$ where the ECC problem is more frequently triggered. 107

1

Introduction

> *"The fundamental observation is that neuroscientists are interested in a particular task or behavioral capacity, and the functional contribution of components of the system to the task of interest. For this reason, we want our model to be able to actually perform that task."*
>
> – Cao and Yamins, 2021

1.1 Motivation

Intuitive physics, the knowledge of the outcomes of physical interactions between macroscopic objects, has been used as the backdrop to study fundamental questions about learning in humans. Intuitive physical knowledge is obviously seen in mature human adults, but also in early infancy (Spelke et al., 1992). A variety of mechanisms have been hypothesized to underpin physical intuition. On the one hand, Nativist accounts (Keil, 1979; Spelke et al., 1992, 1994b; Baillargeon, 2008; Carey and Spelke, 1996; Carey, 2000) e.g. the well-known "core knowledge" hypothesis (Spelke et al., 1992, 1994b) have argued for domain-specific constraints on development. In this framing, some unspecified mechanism in evolution imparts domain-specific cognitive

constraints on an infant's interpretation of experiences. Alternatively, Empiricist views argue for domain-general principles of learning (McClelland, 1988; Mandler, 1992; Elman et al., 1998; Shultz et al., 1994). The forms of these mechanisms paint vastly different pictures of the origins of human knowledge.

Nativist-empiricist debates have raged for a long time in various domains. Why should intuitive physics be particularly diagnostic on this topic? The appetite to understand physical intuition stems partially from the domain's ubiquity. All physical behavior invokes concepts from intuitive physics; it is inescapable. As such, the domain of intuitive physics dwarfs most other domains in terms of how much experience we have with it, both within an individual's lifetime and across our evolution as a species. This fact alone lends support to Nativist accounts. If evolution were to bake in any domain-specific constraints, it stands to reason it would choose a ubiquitous (and relatively unchanging) domain like intuitive physics. This same fact also lends support to Empiricist accounts. If domain-general statistical learning principles could account for human behavior, surely it would be in a domain with an abundance of experience from which to learn. We can see here how understanding the origins of intuitive physics carries serious potential to understand the origins of human knowledge.

The early emergence of physical intuition forms another reason to study it as an indicator of the origins of human knowledge. Let's imagine the dream outcome for the Nativist narrative: the ability for physical reasoning is observed in infants *immediately* at the time of birth. Nativists appropriately conclude that physical reasoning is enabled by innate ideas since there was no time to learn from experience (excluding arguments about experience in the womb for the sake of simplicity). Empiricists would be forced to agree. This would shed light on the origins of knowledge: it would be evidence that evolution can provide "conceptual" constraints. Now what if physical reasoning was shown just a minute after birth instead of immediately after birth?

Nativists would likely draw the same conclusion, but have to make a brief argument why it's unlikely that one minute of experience was sufficient to learn physical reasoning from experience. What if the interval became one day or one week instead of just a minute? As the infant acquires more and more experience before demonstrating physical knowledge, it becomes harder and harder to dismiss the possibility that this knowledge emerged purely from experience and not innate constraints. Even before the work of Baillargeon et al. (Baillargeon et al., 1985b), physical intuition was seen as a likely candidate for "early" knowledge (Gibson, 1979): knowledge that would be detectable while minimizing the explanatory poison of additional experience.

As it stands, nativists have in fact pursued the dream of establishing physical intuition in very young infants, limiting the amount of experience which could be used against nativist claims. Despite citing constraints on the development of attention and visual perception which may pose hard limits on how early physical reasoning can be demonstrated using current methods (Spelke et al., 1992), developmental psychologists have produced strong evidence of physical reasoning in infants as young as 2.5 months of age. Over more than a decade of research they have built a map of how physical intuition develops with age across a variety of physical principles and scenarios. By showing that the physical intuitions which are detected earliest persist throughout development into adulthood, nativists have advocated for the "core knowledge" thesis (Spelke et al., 1994b). This is the notion that the principles of object continuity and solidity [1] constrain infant reasoning about visual experience and form a core upon which mature belief builds [2].

Modern empiricists, specifically connectionists like (McClelland, 1988; Elman et al., 1998; Shultz et al., 1994), have built their own line of supporting research.

[1] Although see (Baillargeon, 2008) for a different hypothesis on domain-specific constraints for physical reasoning i.e. object persistence.

[2] This was a direct rebuke of Piaget's constructivist views wherein infants underwent radical conceptual change as a function of behavioral development. Nonetheless, it staked a strong nativist view that would come into conflict with modern empiricism.

They built artificial systems that used domain-general learning rules and applied them to various domains relevant to the nativist-empiricist dialogue. In particular, (McClelland, 1988; Shultz et al., 1994) applied artificial neural networks to a balance beam task in order to demonstrate that connectionists principles could yield the same stage-like transitions of physical intuition seen in infant development. However, this work was limited in in terms of its scope (weight judgments are just one kind of physical intuition and typically studied in older infants). It was also limited in its fidelity. The balance beam task learned by the artificial neural network was highly simplified.

Ultimately, providing a definitive answer to the origins of knowledge is outside of the scope of this thesis. However, I've described two major accounts (there are *many* others) to highlight a component which both[3] schools lack. As theories of learning physical reasoning, neither view has actually produced a model of that learning. Nativist accounts have posited an initial configuration of the learning problem and described broad attributes on how knowledge expands with development. It's a theory, not a model and as such is limited to coarse-grained predictions. While connectionists have put forth actual models to describe at least one aspect of physical intuition, they have abstracted the task significantly. In doing so, they have sacrificed the ability to model the actual learning problem faced by infants. Its clear that both accounts would benefit from the ability to produce actual models of learning. Furthermore, if the enterprise of understanding the development of intuitive physics is indeed critical to understanding the origins of knowledge *and* probing infant knowledge prior to a certain age is logistically impossible, *modelling* the development of intuitive physics knowledge may be the *only* option we have to understand the origins of knowledge.

It is here where the current thesis makes a concrete and definitive contribution. We develop deep learning models which learn a range of physical concepts from vi-

[3]In fact, I am not aware of any work from any account of the origins of knowledge that provides a faithful account of learning intuitive physics.

sual experience with synthetic three-dimensional objects. To do so, we establish a paradigm for probing intuitive physics knowledge in deep neural networks inspired by the Violation-of-Expectation paradigm. In doing so, we fulfill a prerequisite for any mechanistic model (Cao and Yamins, 2021) of learning intuitive physics: models must be able to actually perform the task they are modelling. The work here lays the groundwork to make both nativist and empiricist accounts of the origins of knowledge more specific and falsifiable.

1.2 Overview of Thesis

My thesis work has investigated the conditions under which deep learning models are capable of learning intuitive physics. In order to do this, we port the Violation-of-Expectation (VOE) paradigm from developmental psychology. We leverage this paradigm to establish a methodology and set of datasets for quantifying physical intuition in deep neural networks. In the rest of this chapter, we give additional background on the scientific context of the current research program. We first give an overview of intuitive physics research in developmental psychology with a special focus on the VOE paradigm. We then describe research from the fields of cognitive science, artificial intelligence, and machine learning which make contact with various aspects of physical intuition. We pay particular attention to the developments in deep learning which made the current thesis possible. Subsequent chapters described the contributions of this thesis. In Chapter 2, we first build a deep neural network architecture with minimal inductive biases and find it only learns physical concepts in a very limited range of experiences. In Chapter 3, we build a model with an object-centric, relational inductive bias which implements domain-specific constraints on physical learning hypothesized by developmental psychologists. We find this model is capable of learning physical principles in fairly general ways and with as little

as 28 hours of visual experience. Furthermore, we find that removing the object-centric inductive bias removes the model's ability to learn intuitive physics. Finally, in Chapter 4 we briefly address how some of the object-centric inductive biases from Chapter 3 could be learned from visual experience.

1.3 Intuitive Physics In Infants

1.3.1 Early Probes of the Infant's Knowledge of Intuitive Physics

Perhaps the earliest experimental work of relevance is Piaget's work on object permanence in infants (Piaget and Cook, 1952, 1954). Piaget found that infants failed to reach for hidden objects until they reached 9 months of age. Given that infants younger than this age can reach for unoccluded objects, this failure to reach must be due to the occlusion. Piaget interpreted the failure to reach as indicative of the infant's conception of the occluded object. Specifically, he believed that infants did not reach for the occluded object because they did not believe the occluded object existed. In short, he believed young infants did not possess object permanence until 9 months of age. He did so on the assumption that infant actions reflect infant knowledge. This was part of an overarching constructivist view of human development. Under this theory, infant's gained knowledge through the interaction between their experiences and ideas. Piaget hoped to demonstrate that, even universally applicable concepts like object permanence which we could expect to be the "earliest" concepts, emerge as a function of experience.

Despite numerous replications of Piaget's experimental results (see (Gratch, 2013) for a review), subsequent research challenged Piaget's *conclusions*. (Spelke et al., 1995) described it succinctly, "Piaget equates having a concept of an object to being able to act on that object. Because Piaget's discussion of object permanence is

couched within his discussion of sensorimotor intelligence, he does not consider perception or cognition in the absence of action." In other words, infants could have the object permanence concept and merely fail to reach for occluded objects due to motor limitations. This hypothesis is partially supported by Piaget's own work showing that 4-8 month old infants fail on "means-end" sequences: they are unable to chain a sequence of actions wherein the preconditions for the second action are achieved by the first action (Piaget and Cook, 1952). An alternative approach, the Violation-of-Expectation (VoE) paradigm, eschewed reach-based tasks in favor of measuring "looking time". The next section describes that methodology and the various ways it has been applied to reveal early infant knowledge.

1.3.2 The Violation-of-Expectation Paradigm

The lineage of the Violation-of-Expectation (VoE) paradigm started with the seminal work of (Baillargeon et al., 1985a). In response to Piaget's claims about object permanence in infancy, researchers were looking for ways to probe object permanence in infancy *without* the requirement of "means-end" sequences. Work by Bower (1971) tested infants in the dark and suggested infants could possess object permanence as early as two months of age. Methodological and theoretical faults (Baillargeon et al., 1985b) prevented the widespread propagation of Bower's approach [4].

Instead, Baillargeon et al. (1985a) developed a probe based on *looking time* which did not require any reaching at all. To test for a specific concept, two visually similar probe displays are shown to infants. One of the probes is physically possible: it is con-

[4] I include this misstep for the benefit of the machine learning and artificial intelligence communities. In those fields, the coin of the realm is to build a model or method to *solve* a task. However, as the example of Bower shows: an equally difficult, and even more important, challenge is to *design* tasks which definitively probe a desired capability. By looking to the cognitive sciences, the field of artificial intelligence can takeaway the following: it should allocate more resources and elevate the importance of task design and it should borrow existing tasks that have made it through the crucible of years or decades of scientific scrutiny. To be fair, a major machine learning conference, NeurIPS, has for the first time this year (2021) introduced the "Datasets and Benchmarks" Track.

sistent with the physical concept being tested. The other probe violates that concept. By measuring gaze duration, which is taken as a proxy for surprise, researchers infer whether or not the impossible display violated the infant's expectations. Thus, in VoE framework, possession of a belief is equated to holding a certain set of expectations.

The experiment of Baillargeon et al. (1985a) was indirect in two ways. First, their specific experimental probe (this is not true for all VoE experiments) of object permanence required knowledge of solidity (the notion that objects cannot interpenetrate each other). Thus, a null result under their experimental design could not distinguish between a failure of object permanence or a failure of object solidity. Baillargeon et al. managed to sidestep this pitfall since they found significant effects [5]. Second, their use of looking time is a more problematic source of indirection (this is an issue for *all* VoE experiments). It is straightforward to conclude that if an infant reaches for an object, they believe the object exists. The link between belief of an object's existence and looking time is less direct. It requires linking looking time to surprise and linking surprise to conceptual knowledge.

Before describing the resolution of this issue, we take a detour to describe the general scheme of the VoE paradigm. Then we overlay the specifics of (Baillargeon et al., 1985a) onto the general VoE structure. Finally, we return to describe how the structure addresses the above concern.

Generally speaking, the VoE paradigm contains an experimental condition and a control condition. Both conditions follow a similar structure: a series of habituation events follow a probe trial. In the experimental condition, this probe trial is either a physically-possible or physically-impossible event. The goal of this condition is to assess whether infants show increased gaze duration at the impossible event over the possible event. Both events are as visually similar as possible, but, by necessity, cannot be the same exact display. In the control condition, the probe trials are both

[5]Despite being inconsequential for their results, we note it here to highlight the fact that probes of intuitive physics can sometimes depend on more than one physical concept.

physically possible, but match aspects of the experimental probe trials. The goal of this condition is to assess whether infants look longer at features of the physically impossible probe that does not violate any physical concepts. This is used to rule out the explanation that increased gaze for impossible events is simply due to low-level visual features of the physically impossible display and not at all related to conceptual understanding. Individual infants are assigned to either the experimental condition or the control condition and are exposed to the probe trials in alternating sequence.

Baillargeon et al. (1985a) developed and instantiated the schema above to probe infant knowledge of object permanence in five-month olds. The centerpiece of their experimental design was a drawbridge [6]. This drawbridge started flat on the display surface and rotated a full 180° away from the infant until it once again was flat on the display surface. It then rotated back in the opposite direction to its starting position. This constituted the habituation event in both the experimental and control conditions. The probes in the experimental condition added a box that was placed in the path of the drawbridge. Importantly, the box was fully occluded by the drawbridge at a 90° rotation. In the possible event, the drawbridge rotated past the 90° mark to 112°. At that point, it rested against the box and eventually rotated back down to its initial position. In the impossible event, the drawbridge rotated a full 180° seeming as if it went through the box (which should have prevented its motion). It then rotated back to its original position, revealing the box in its original location. The logic is that if objects are solid *and* continue to exist even when occluded, then the box should have prevented the drawbridge from rotating all the way. When infants are surprised by this impossible display, it is taken as evidence that they hold the logic above about the concepts of solidity and object permanence[7]. In the control condition

[6]Some researchers even refer to the VoE paradigm as the "drawbridge paradigm" (Munakata, 2000; Rivera et al., 1999)

[7]There are, of course, physically possible mechanisms for generating the event shown to the infant. After all, these displays were not computerized. It remains a humorous possibility that infants look longer at the "impossible" display simply because they are in disbelief that someone would go through all of the trouble of making such an illusion. More importantly, I flag this as a

(motivated below), the box is placed outside of the range of the drawbridge (either next to or behind it) to allow the drawbridge to freely rotate without any "magic tricks" as in the experimental condition. As in the control condition, the drawbridge rotates with 112° or 180°.

We can now address how the VoE paradigm draws conclusions based on differences in looking times. A critical finding for the VoE paradigm is that looking times can be used as an indication of novelty (Fantz, 1964) in infants. This novelty signal is then interpreted as a proxy for surprise (i.e. the violation of an expectation) in infants. Given this premise, the question turns to which aspects of the impossible display does the infant find surprising relative to the possible display? There are two salient differences between the displays. First, there is the (putatively) conceptual difference: the impossible physics appears to violate the principles of solidity or object permanence. Second, there is a difference in the visual statistics of the two displays. In the impossible display, the drawbridge rotates 180° while in the possible display it only rotates 112°. This second difference between the displays motivates the control condition. It is used to rule out that infants are somehow more surprised (or just have a general looking preference) for the full rotation seen in the impossible display. Because the results of Baillargeon showed no difference in looking times between the two control displays, it is concluded that infants are surprised by the only remaining difference between the two experimental probes: the (seeming) violation of a physical principle.

---

hypothesis over infant beliefs that is immediately discarded due to being judged too implausible. I have to assume (no one has explicitly addressed this issue) that the underlying logic is that infants lack either experience or the inference mechanisms to have a non-zero belief or awareness that the impossible display was an elaborate hoax constructed for their benefit. In short, researchers have dismissed a whole set of alternative hypotheses about why infants look longer at impossible displays if those hypotheses invoke experiences or mechanisms presumed too complex for infants. In regards to our work on learning intuitive physics and the virtual impossible displays we constructed, we found reviewers were less likely to automatically dismiss hypotheses that invoked complex mechanisms outside of the domain of experience of our models.

The VoE method was subsequently deployed in a myriad of ways to answer various questions. The pursuit of those questions lead to a timeline of the visibility of infant physical intuition of different concepts. We survey some of them to highlight the kinds of questions that can be asked with the VoE paradigm. Furthermore, the centrality of intuitive physics to understanding learning. Having applied the VoE paradigm to 5-month olds, (Baillargeon, 1987) wondered if physical intuition emerged even earlier as a more potent rebuke to the knowledge in action hypothesis. Indeed, Baillergeon et al. established that 4.5-month olds and some 3.5-month olds ("fast habituators") gazed longer at violations of object permanence than their physically-plausible counterparts. (Spelke et al., 1992) argued that infants do not undergo radical conceptual change as a function of perceptual or behavioral development and instead build up their beliefs around "core knowledge." They found that 2.5 month old infants could reason about the principles of continuity and solidity, but not gravity and inertia. By making an appeal about the centrality of continuity and solidity to mature physical reasoning, they concluded these results were consistent with the core knowledge hypothesis. Spelke et al. (1994a) extended that work and found that weak reactions to violations of inertia emerged in 8- and 10- month olds but not 4- or 6-month olds. Needham and Baillargeon (1993) developed an alternative VoE paradigm to probe gravity, which they referred to as support. They tested 4.5-month olds intuitions about support and found, in contrast to Spelke's findings, that they *were* surprised when blocks hovered in mid-air.

Whereas the aforementioned previous experiments generated variants of the VoE paradigm through application to different physical concepts, a body of work (Baillargeon and DeVos, 1991; Aguiar and Baillargeon, 1999) emerged which generated variations within the same physical event. Instead of asking whether or not an infant was sensitive to a physical principle in a yes or no fashion, they investigated which object properties and relations lead to successful discrimination of possible versus

impossible displays by the infants. This is a way to probe what *variables* are used in physical judgments by infants. Throughout several studies infants were shown a toy moving behind two occluding pillars. In the physically possible displays, the toy was initially visible, occluded by the first pillar, visible again in the space between the two pillars, occluded by the second pillar and finally emerged from behind the second pillar. In various conditions, the violations of continuity happened when the toy was magically occluded when it should have appeared between the two pillars. It appeared to magically teleport from one end of the display to the other. At 2.5 months old, infants were surprised when the toy failed to appear between two unconnected pillars. However, this effect went away if a thin strip was placed connecting the two tops of the pillars (there was still a large visible gap between the pillars where the toy should have appeared). This suggests infants had a very simple "behind/not-behind" variable that did not take into account the relative heights of the occluder and moving object. By 3 months of age, infants are sensitive to the placement of this strip. If the two pillars are connected at their bases with a strip, then they are again surprised by the violation. By 3.5 months of age, infants become sensitive to the relative height of the occluder and toy. This was extended to other object properties (e.g. by 7.5 months infants realize occluders must be opaque). (Baillargeon et al., 1992) also applied this approach to probe which relational properties affect infant perception of when an object should be supported by an object underneath it. This approach of mapping the emergence of variables was itself extended. (Hespos and Baillargeon, 2001a) again investigated which variables were used in infant physical reasoning. However, they wanted to assess whether a particular variable, once acquired, was used in *all* physical judgments. They contrasted VoE results for occlusion events against containment events and found that infants used the "relative height" variable for occlusion, but not containment events. This suggested a category-specific

attention to object properties and interactions (which formed a prominent aspect of Baillargeon's account of infant intuitive physics: the "physical reasoning system").

1.4 Intuitive Physics And Artificial Systems

1.4.1 Learning Physics

Investigating intuitive physics in artificial systems has seen an increased interest due to advances in deep learning. However, it is not a new endeavor. As early as (McClelland, 1988), cognitive scientists used artificial neural networks to understand learning of physical principles. McClelland et al. modelled a balance-beam task used by (Inhelder and Piaget, 1958) where weights were placed on a balance beam and the task was to determine which way it would tip. He argued that the application of a simple learning rule to connectionist networks, implemented by the back-propagation procedure (Rumelhart et al., 1986), could describe children's stage-like progression through rule-like behaviors seen in children's balance judgments. This approach was in contrast with competing accounts that saw rule-like behavior and concluded they must be the product of an explicit rule-based system (e.g. (Klahr and Siegler, 1978)). Despite the "drastic oversimplification of the human mind and of the task" in the connectionist system, it's now clear that McClelland's work compelled the field to take a connectionist account of development seriously. In that vein, the work of (Shultz et al., 1994) modelled the balance beam task with a cascade-correlation network (Fahlman and Lebiere, 1990) and argued it was more general than McClelland's specialized architecture and better captured developmental behavior. Although again they used an impoverished version of the balance beam task which completely abstracted away perception and action. Battaglia et al. (2013) used a ready-made physics engine as a forward model to perform a number of classification and regression tasks related to the dynamics of a tower of falling blocks. (Lerer et al., 2016; Zheng et al., 2015)

studied the use of convolutional neural networks for predicting block tower stability from image data. (Battaglia et al., 2016; Chang et al., 2017a; Watters et al., 2017; Battaglia et al., 2018) are emblematic of a wave of deep learning research which advocated for a relational inductive bias in the service of learning the dynamics of physical systems such as 2D bouncing balls and mass-spring systems. Subsequent work pushed this framework using *graph neural networks* to simulate complex physics entailing thousands of particles (Sanchez-Gonzalez et al., 2020) and mesh simulations (Pfaff et al., 2020). The maturation of graph neural networks and the general object-central, relational inductive bias they support form a critical building block for our work in Chapter 3. Additionally, there is a body of work that emerged concurrently to the research in this thesis which directly aims to assess intuitive physical knowledge in artificial systems. We discuss that work in Section 3.3.

1.4.2 Discovering Objects

As described above, recent years have seen an increased use of object-centric inductive biases in deep learning models. These are models which reason and compute over objects (which can be more generally thought of as discrete entities or slots). By adopting this structure, models gain the power to generalize in regimes that would normally not be supported by their training data. The cost to adopting this structure is that it incurs explanatory debt. If a model requires object-level information, either as exact object state or via segmentation masks of visual input, the question naturally arises: where do these object representations come from? Without providing a mechanism by which object representations can be learned from experience, the parts of the model which *are* learned through experience are undermined. This is both a practical and theoretical concern. Practically speaking, requirements on labelling parts of the inputs to a system (e.g. manually annotating the objects in a visual sensor) to a system limit the applicability of that system. Theoretically speaking,

we know humans do not get dense visual annotation of object boundaries. We know that humans generalize to completely new object shapes and textures. A system that can predict the dynamics of the physical world only when given object segmentation masks is of obviously limited utility. As such, learning to factorize the world into objects has been a longstanding task in the fields of machine learning and computer vision.

Work in machine learning on parsing objects falls into two broad categories. One class of approaches, object detection, uses supervised training data. As this approach offers little explanatory power for the current work, we review it only briefly. Approaches in this regime, require labelled training data either in the form of "bounding boxes" around objects (e.g. (He et al., 2017; Carion et al., 2020)) or dense per-pixel annotations (e.g. (Lin et al., 2017)). However, they can generalize to new objects. As such, they provide utility when the goal is simply to get an object representation that can generalize to new data. This means it would not be necessary to label all possible objects a system might encounter, substantially reducing the costs and effort of deploying such a system into complex environment. In fact, these models have shown impressive performance on real-world images. From an explanatory perspective, however, these supervised approaches do not provide an account of plausible mechanisms for learning to partition the world into objects.

Recent years have seen an increase in *unsupervised* approaches to object decomposition (sometimes called object discovery). These approaches rely on various principles and heuristics to factorize an image, or sometimes sequence of images, into discrete entities. We begin with the most relevant work in this area, not the earliest. As this area is rapidly expanding, we highlight representative developments instead of exhaustively covering all related works. (Burgess et al., 2019) build a model, MONet, which learns a set of attention masks and independently reconstructs the pixels belonging to each mask using a variational auto-encoder. The rationale is that scenes can be de-

composed into a set of abstract motifs or objects. Assuming this structure allows for a much more efficient representation of a scene dataset. Instead of memorizing all the pixels in all the images, each image can be described by a list of the objects that occur in that image. Because the objects recur across multiple scenes (either exactly or in basic variants), each object can also be represented more compactly than simply listing the constituent pixels in that object. The work of (Burgess et al., 2019) is to build a system that is incentivised towards this specific type of economy of representation (i.e. both at the object-level and the scene-level). MONet was successfully applied to two different synthetic datasets composed of 3D shapes. Beyond simply segmenting the image into objects, Burgess et al. showed that the object representations learned by MONet in-painted regions of the object that were occluded. An earlier version of this idea, Attend-Infer-Repeat (AIR) (Eslami et al., 2016), also factorizes images into a set of objects. AIR hard-codes the representational variables used for objects: each object has 'what', 'where', and 'presence' variables. This increases interpretability of the model's representations. AIR, however, failed to model partially occluded objects and was incapable of modelling backgrounds. SPACE (Lin et al., 2020) is a model with explicit variables like AIR but adds parallel spatial attention and a dedicated pathway for parsing the background. This model is highly scalable (theoretically a minimum of up to $24 * 24$ objects). They demonstrated results on a synthetic "3D-Rooms" dataset and Atari games. In practice, we found it was quite difficult to use this model in domains where objects had varying sizes.

Sequential Attend, Infer, Repeat (SQUAIR) (Kosiorek et al., 2018) extended the approach of AIR by attending sequentially over frames in a video, thus improving the model to recognize partially overlapping objects. However, these results were limited to binarized images. (Greff et al., 2019) built a model, IODINE, which iteratively refines an initially-random factorization of an image into objects. Whereas SQUAIR iterated over frames in an image, IODINE iterates over a single image. The refinement

iterations in IODINE can be thought of as inter-object arbitration. An initial factorization may split a single object across two object slots. The refinement steps "shift" split objects into a single slot and accrue details. This approach yielded convincing object decompositions on the CLEVR dataset (Johnson et al., 2017) of synthetic 3D objects. An intriguing aspect of this model is its behavior when applied to ambiguous Tetris configuration. Two Tetris pieces of the same color were placed next to each other. The configuration was chosen so that there were two possible arrangements of Tetris pieces that would create the image shown to the model. Running IODINE with different initially-random factorizations showed that it was capable of correctly factorizing the same input into a variety of valid Tetris pieces. Additionally, IODINE is quite easily extended to track objects over videos (without complete occlusions), by refining for a number of steps within an image in a video and then refining for more steps on the next video frame and so on. The only major drawback of this method is that it is quite resource-intensive to train. The work of (Locatello et al., 2020) introduced the Slot Attention module (SAM) which could be applied to object discovery. SAM iteratively attends over an image to populate a set of slots which can be used for various downstream computations. When that downstream task is image reconstruction, SAM yields an object decomposition architecture. This work again showed compelling unsupervised object decomposition on the CLEVR dataset. It is less resource-intensive than IODINE, while maintaining similar benefits (iterative refinement).

2

Learning Intuitive Physics Without Objects

My first project had two goals. First, it wanted to establish a methodology for probing intuitive physics knowledge in artificial systems. Second, it sought to push the limits of an empiricist account of intuitive physics: can intuitive physics, given recent advances in deep neural networks, be learned from vision without domain-specific constraints (i.e. object-centric computation) on learning? To do this, we made a synthetic dataset which probed various physical concepts using the Violation-of-Expectation (VoE) paradigm. We applied a then state-of-the-art neural network, a variational recurrent neural network, to our dataset. Importantly, we did not endow this network with an architecture that explicitly promoted object-centric representation and computation, as nativists would suggest are the appropriate constraints for learning intuitive physics. We found that although it was capable of learning intuitive physics, this was only possible in a very narrow sense. The network did not generalize its experiences from arbitrary physical events to the probes of specific physical concepts in our "virtual laboratory". This work firmly cemented the VoE for artificial systems paradigm. Despite caveats about its generalization abilities, this

was the first artificial system that could be said to learn a broad range of intuitive physical principles from perceptual data, where learning was defined in a systematic and quantifiable way. This project was posted to the pre-print server *arXiv* and this chapter is a reproduction of that article. The full citation for the paper is:

Luis Piloto, Ari Weinstein, Dhruva TB, Arun Ahuja, Mehdi Mirza, Greg Wayne, David Amos, Chia-Chun Hung, and Matthew Botvinick. Probing physics knowledge using tools from developmental psychology. *arxiv*, abs/1804.01128, 2018. URL http://arxiv.org/abs/1804.01128

Probing physics knowledge using tools from developmental psychology

2.1 Abstract

In order to build agents with a rich understanding of their environment, one key objective is to endow them with a grasp of intuitive physics; an ability to reason about three-dimensional objects, their dynamic interactions, and responses to forces. While some work on this problem has taken the approach of building in components such as ready-made physics engines, other research aims to extract general physical concepts directly from sensory data. In the latter case, one challenge that arises is evaluating the learning system. Research on intuitive physics knowledge in children has long employed a violation of expectations (VOE) method to assess children's mastery of specific physical concepts. We take the novel step of applying this method to artificial learning systems. In addition to introducing the VOE technique, we describe a set of probe datasets inspired by classic test stimuli from developmental psychology. We test a baseline deep learning system on this battery, as well as on a physics learning dataset recently posed by another research group. Our results show how the VOE technique may provide a useful tool for tracking physics knowledge in future research.

2.2 Introduction

The goal of developing artificial agents that display an intuitive grasp of everyday physics – a basic understanding of the behavior of objects and forces – has been widely acknowledged as a core challenge in artificial intelligence (AI) (Denil et al., 2017; Ullman et al., 2017). In many settings, the aim is to deploy artificial agents in physical

environments, whether real or simulated. In such contexts, physics understanding is of obvious importance, not only because it supports effective interaction with objects, but also because it supports sample-efficient learning and broad transfer (Wu et al., 2015; Finn and Levine, 2017). For related reasons, research in cognitive science has flagged intuitive physics as a 'core domain' of conceptual knowledge (Spelke and Kinzler, 2007), emerging early in life (Spelke, 1994) and providing the foundation for concepts in many other domains (Bremner et al., 2015; Hespos and vanMarle, 2012).

Research on intuitive physics has been divided between two strategies. In some work, the approach has been to directly leverage physics knowledge, for example by integrating a ready-made physics engine into the agent architecture (Wu et al., 2017). In other work, the goal has instead been to learn physics from raw sensor data (Denil et al., 2017; Botvinick et al., 2017). The latter case – our focus in the present work – carries a special challenge, which is to evaluate what knowledge the agent has acquired. When the goal is to learn physics, it is essential to have some method of tracking the course of learning and measuring its success.

One widely applied method for assessing physics knowledge is to train systems to produce explicit predictions concerning the evolution of physical systems (Lerer et al., 2016; Battaglia et al., 2016; Chang et al., 2017a), and to use the accuracy of such predictions as a measure of physics knowledge. While this technique can be quite powerful, it must be applied with delicacy, since both successes and failures of prediction can be tricky to interpret. Inaccurate predictions can derive from quantitative (e.g., metrical) errors rather than flawed conceptual knowledge. For instance, mean-squared error in pixel-space would penalize inaccuracies in predictions for large objects more than small objects, but this would not reflect the importance of size in physical *concepts*. Conversely, successful prediction generally relies jointly on a set of physical concepts, making it difficult to crisply analyze the content of acquired knowledge.

In the present work we introduce and evaluate a complementary method for probing physics knowledge in artificial systems. Our approach is directly inspired by research in developmental psychology. In particular, we adopt two key ideas. The first is to assess physics knowledge by targeting specific physical concepts. Developmental psychology has, over the past fifty years, identified a core set of principles, which together provide a foundation for intuitive physics. These include concepts such as 'object persistence', the fact that objects (even when occluded) do not wink in or out of existence; 'unchangeableness', that size, shape, pattern, color do not spontaneously change; 'continuity', that moving objects, unless perturbed, will follow smooth trajectories; 'solidity', that solid objects cannot interpenetrate; and 'containment', that objects remain in their containers when containers are moved (Baillargeon and Carey, 2012a).

In addition to isolating these basic principles, developmental psychology has also invented and refined what is by now a widely accepted and replicated experimental technique for probing their acquisition, referred to as the *violation of expectations* (VOE) method (Baillargeon et al., 1985a). Here, children are presented with dynamic physical displays, some of which violate a basic physical principle, essentially taking the form of 'magic tricks' in which objects disappear, float unsupported, or otherwise contravene physical law. The critical measure is surprise, typically quantified in terms of looking time. When a child stares longer at a display that violates a physical principle than a carefully matched display that does not, this provides evidence that the child understands the relevant principle (Ball, 1973).

In the present work, we introduce the VOE method as a tool for assessing intuitive physics understanding in AI systems. In particular, we make three contributions. First, we formalize the technique, leveraging psychological notions of surprise that afford a direct quantitative link with AI methods. Second, we introduce a collection of procedurally generated datasets, closely modeled on work in developmental psychol-

ogy. Finally, we apply the VOE method to a natural baseline model, establishing a benchmark for next-step research. Our results indicate that systems with intuitive physics knowledge may not be far out of reach, but that important challenges nonetheless remain.

2.3 The Violation of Expectation Paradigm

As noted above, in developmental psychology the dependent measure usually employed in VOE studies is looking time, which is well established to correlate with subjective surprise. Itti and Baldi (2010) demonstrate a link between gaze time, and gaze direction to a specific computational measure of surprise. In particular, they showed that experimentally observed looking times are well predicted by the estimated Kullback-Leibler divergence (KL) between prior expectations and posterior beliefs given perceptual inputs. This was further reinforced in subsequent work by Téglás et al. (2011).

Our novel contribution is to capitalize on this preceding work by applying the same measure to AI systems. Specifically, we look at the KL between the latent prior and posterior distributions in suitably designed systems, and treat this as developmental psychologists treat looking times. This approach requires that the AI system under consideration supports calculation of the KL. Fortunately, the literature provides a general class of models that have this property, based on deep variational inference (discussed further in section 2.5).

2.4 Dataset and Probe Design

In order to apply the VOE method to AI systems, we developed an initial collection of datasets, taking direct inspiration from experiments in developmental psychology. This corpus is described in general terms, and we then provide details of five spe-

cific datasets. The full corpus will be made available online (for sample videos, see https://goo.gl/M1Heho).

2.4.1 General Dataset Design

The datasets we introduce consist of videos that are procedurally generated and run in the Mujoco physics engine (Todorov et al., 2012). Videos cover approximately 3 seconds of simulated time and are created in categories focusing on specific intuitive physics concepts, with each category separated into testing, validation, and training sets. Test and validation corpora comprise paired *consistent* (plausible) and *inconsistent* (implausible) probes. The training corpus for each category consists of *examples* and *controls*, each discussed below.

Paired probes form the basis of experimentation in the VOE paradigm (see Figure 2.1). Both paired probes are identical aside from a manipulation in the inconsistent probe which makes the scene physically implausible, and is carefully controlled to isolate particular concepts in intuitive physics. In order to ensure the system is reasoning over the history of the scene (and not, for example, simply encoding the statistics of single frames), we ensure that the manipulation is never immediately observable. This is usually accomplished by having the manipulation occur while occluded.

In addition to test probes, we produce a training corpus which is made of *consistent examples* as well as *controls*. The consistent examples are similar to the consistent probes but have a much greater degree of variability (see Figure 2.2). Controls are designed to mitigate potential biases in the dataset, as illustrated through specific examples below.

Choices made in designing the train and test data were guided by three desiderata: diversity, control/matching, and train-test separation. Diversity includes all forms of variability that are introduced in the datasets. These include altering object shape,

size, pose, event timing, floor pattern, and any other aspect selected during procedural generation. Often, variability introduced in consistent training examples is so large that constraints imposed on the probe data are overstepped (for example, an intended occluder may not always occlude a target object in train data). Variability was maximized among probes as well, but subject to the constraint that the resulting video must demonstrate the phenomenon of interest (or violate it, in cases of inconsistent probes)

The addition of controls ensures there is training data that matches the inconsistent probes after the manipulation. This, however, is done in a manner that is consistent with physics (so generally controls begin differently from inconsistent probes but end similarly). Because of the inclusion of controls in the training data, the model should not be surprised by an inconsistent probe simply because the scene ends in a manner different from consistent examples, and instead be surprised by the manipulation due to the context established earlier in the video.

Additionally, we counter-balance probes to ensure that the model cannot be surprised by some superficial aspect of the probe data (for an example, see section 2.4.2). This means each dataset contains two types of consistent and inconsistent probe distributions.

To summarize, our dataset is characterized by different distributions that produce train and consistent examples, as well as carefully considered controls that push the model to recognize the difference between scenes that can be superficially similar but semantically quite different. In combination they represent a concerted effort to introduce significant train/test separation and require rich generalization by the model.

We now describe key aspects of the specific datasets employed in our experiments.

2.4.2 'Object Persistence'

Perhaps the most fundamental aspect of intuitive physics is understanding that objects cannot disappear from existence, and is often called 'object persistence' (or 'permanence'). Taking inspiration from a classic behavioral experiment (Baillargeon et al., 1985a), probes for this category involve a rigid plank falling on an object. In the consistent probe, when the plank falls on the object, the plank occludes it while also remaining propped up by it as expected (Figure 2.1*a*, top row). By contrast, in the inconsistent probe the plank falls on top of the object (in a manner that is initially identical) but ends up flat on the floor, as if the object had disappeared (Figure 2.1*a*, bottom row). In the counterbalanced probes, the consistent probe has the plank falling flat on the floor in an otherwise empty scene, and the inconsistent probe has the plank falling in the same empty scene but ends up inexplicably propped up by an item that is made to appear under the plank while it occludes part of the floor.

Variability in the consistent examples arises from using an object which is either a sphere or cube, and of differing sizes. The width and height of the plank is allowed to vary greatly as well. Furthermore, both items have randomized initial poses so that in each video the plank falls in a manner that is unique and difficult to predict (Figure 2.2*a*, top row).

In the controls, we must ensure the system does not learn simply to associate the initial presence of a cube (or sphere) with the plank remaining supported, so controls consist of videos where the plank falls on an open location in the floor while the additional object is visible but does not support the plank (Figure 2.2*a*, bottom row). Another control has the plank falling on the floor with no other object present, perceptually matching the inconsistent probes after the manipulation.

2.4.3 'Unchangeableness'

By the principle of 'Unchangeableness' (Baillargeon and Carey, 2012a), objects tend to retain their features (e.g., color, shape) over time. In the consistent probes of this dataset, a random assortment of static objects are aligned in the foreground (Figure 2.1*b*, top row). A screen is lowered in front of those objects, and is then raised. The training examples unfold in the same manner, but objects may be scattered anywhere in the scene, and the screen has more variable size and may appear anywhere from the foreground to the background (Figure 2.2*b*). This means that only some (or none) of the objects may be occluded, and may be occluded at different times depending on the vertical position of the screen. Controls resemble training examples with either objects or the screen omitted.

The concept of unchangeableness relates to a number of different aspects of objects, and therefore we include multiple inconsistent probes. The first set of inconsistent probes examines the unchangeableness of position of static objects, and moves the position of one or more of the objects while occluded. The other two inconsistent probes manipulate either the color or shape (see Figure 2.1*b*, bottom row) of one or more objects while occluded.

2.4.4 'Continuity'

The concept that an object traces out one continuous path through space and time is referred to in the developmental literature as 'continuity' (Spelke, 1994). For videos in this category we use a nearly identical setup to a classic experiment (Aguiar and Baillargeon, 1999) where consistent probes begin with two static pillars separated by a gap (Figure 2.1*c*, top row). A ball is rolled horizontally behind both pillars so that it is visible before, between, and after the pillars during its trajectory. In the inconsistent probes, while the ball is occluded by the first pillar, the ball is made

invisible for the period when it would be between the pillars, and then reappears after the second pillar (Figure 2.1*c*, bottom row).

Consistent examples (Figure 2.2*c*, top row) are equivalent to consistent probes, but the ball is not constrained to be on a trajectory that is occluded by the pillars. In one form of control (Figure 2.2*c*, bottom row) there is an additional large occluding block that fills the gap between the pillars, resulting in the ball being visible in only the same locations as in an inconsistent probe.

2.4.5 'Solidity'

This dataset is a recreation of an experiment that uses an object and an occluder (Hespos and Baillargeon, 2001c) to test understanding of the solidity of objects, as related to the penetration of an object through a container and the ground below. In probes, perspective is carefully controlled such that the camera can view inside the top of the container but not the bottom. In consistent probes (Figure 2.1*d*, top row) a rectangular block is dropped into the container and comes to rest as expected. In the inconsistent probes (Figure 2.1*d*, bottom row), the object "falls through" the container and the floor, and therefore disappears from view (with the penetration itself occluded by the face of the container) even if the object should remain visible due to its height. Consistent examples (Figure 2.2*d*, top row) are similar to the probes but do not control for perspective; objects range from short enough to be completely occluded by the edge of the container to tall enough to protrude from the top of the container. Controls include scenes where no object is visible, which perceptually matches the inconsistent probes after the object falls out of view, as well as scenes where the object falls behind the container (Figure 2.2*d* bottom row, so that it is occluded by the container, but in a different manner).

2.4.6 'Containment'

Finally, we look at containment events, which requires understanding that an object inside a container remains inside as the container is moved. As is the case with earlier studies in developmental psychology (Hespos and Baillargeon, 2001b), probes of this dataset involve an object falling inside a container. Once the object (a cube) falls to the bottom of the container it is occluded by the container's walls, and the container is then moved by a rod that descends from the top of the scene (Figure 2.1*e*, top row). In the inconsistent probes (Figure 2.1*e*, bottom row) the movement of the container reveals the cube to have remained in the same position on the floor (where it and the container previously were), and consistent probes have the cube moving inside the container as expected. Consistent examples (Figure 2.2*e*, top row) are like consistent probes, but container position and size are not constrained to occlude the cube. In controls (Figure 2.2*e*, bottom row) the cube may fall behind the container (therefore also becoming occluded), and when the container is moved the cube remains where it was (similar to inconsistent probes). The cube may also fall at a completely randomized position before the container is moved.

2.5 Model

In order to make an initial test of the VOE method, we require a model that has some form of memory, and readily allows for computation of the KL-divergence. Alhtough we could have selected from a number of models that meet this criteria, we chose a previously proposed generative temporal model with external memory (Gemici et al., 2017), as it showed good initial performance. This takes the form of a variational recurrent neural network (VRNN) with the Least-Recently Used (LRU) memory mechanism for memory. This is most easily understood as a recurrent neural network where each time step or "core" is a variational autoencoder (VAE) with the

hidden state determined from the external memory. Unless where stated otherwise, implementation details are exactly as in the LRU model of Gemici et al. (2017). One important deviation from standard variational autoencoders is that in the case of VRNNs, the prior distribution is generated at each time step as a function of the previous observations in the sequence.

The model operates on a sequence of observations, $x_{\leq T} = \{x_1, x_2, \cdots, x_T\}$ and approximately infers a set of latent variables $z_{\leq T} = \{z_1, z_2, \cdots, z_T\}$. In order to specify our VRNN, we must specify four maps as defined in (Gemici et al., 2017). The *posterior map* defines the approximate posterior distribution over 256 Gaussian latents. This is calculated by adding to the prior the output of a multi-layered perceptron (MLP) with 2 layers of 512 units using the *tanh* activation function clipped between $[-10, 10]$. The MLP takes as input an image passed through a convolutional neural network (described in the Appendix), the memory output determined at the previous timestep, and another copy of the prior for the current timestep. The *prior map* specifies the prior distribution over latents using the same architecture as the posterior map (with its own set of weights, of course). It depends only on the *history* of latent variables and observations via the external memory's output. This memory output is the same output fed to the posterior. The *observation map* specifies the parameters of the likelihood function via a deconvolutional neural network (see Section A in Appendix) as a function of the sample from the posterior distribution. The *transition map* specifies how the hidden state, h_t, and memory, M_t, are updated at each time step. For this we use the LRU memory mechanism with a long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997a) controller using 128 units to control reads. Output from the memory is read as the weighted sum over 3 memory slots. The LSTM takes as input the previous hidden state and the sampled latents inferred for the previous observation.

The external memory has a total of 15 memory slots and 100 units per slot. We use a Bernoulli distribution for the likelihood of our observations by scaling pixels to be in the range $[0, 1]$. We also include an additional term in the objective function to regularize the prior distribution (and by proxy the posterior distribution), towards a unit Gaussian. This is scaled by $1e^{-4}$ in the objective function.

2.5.1 Training and Evaluation

For each concept-category dataset, we trained on 100,000 sets of consistent examples and controls each. Testing was conducted over 2,000 pairs of probes for a hyperparameter search, and another 8,000 pairs of probes were used for validation after the hyperparameter sweep was completed. Each video was restricted to 15 frames at 64x64 resolution in RGB channels. Training was done under two regimes. In the first case, we trained and tested separate models each on a single dataset. In the second case, we trained on all the datasets together. In all cases, we used the ADAM (Kingma and Ba, 2015) optimization algorithm with a batch size of 10, trained for 2 million steps, and a learning rate of $5e^{-6}$. Although the loss function converged to a stable value well before this point, we noticed continual improvement in reconstructions as well as reductions in relative surprise with continued training in some categories (see Figure 2.7).

As the probes were evaluated by the model, the KL-divergence was recorded over all frames for both the consistent and inconsistent probes. Statistical testing for the KL-divergence between the consistent and inconsistent probes was done with a one-tailed, paired t-test with a significance level of 0.05.

2.6 Results

We first evaluated our baseline model trained on each dataset separately. For all categories we found the surprise for inconsistent probes was higher than consistent probes on average (see Figure 2.4, red text for exact numbers). This indicates the network does have an ability to recognize violations of these concepts, although the trial-by-trial variability in the histograms in Figure 2.4 indicate there is room for model improvement.

Although our focus in the present work is on the VOE method, it is worth noting convergent evidence that the network was "understanding" scene information in a meaningful way. Visualizations confirm the network attempts to reconcile its knowledge of physics with the violations that occur during inconsistent probes. Figure 2.6 presents an Object Persistence inconsistent probe (top row) along with renderings of the prior per frame (bottom row). In this video, a magenta cube disappears while occluded by a falling cyan plank. Here the prior, conditioned on the preceding frames, predicts the sheet should rest on the magenta cube, instead of having the cube disappear. This is apparent in the portion of the magenta cube protruding beyond the top of the plank in the last four frames.

While we did not explore generalization beyond the training distribution systematically, initial experimentation indicated that the network was capable of some degree of extrapolation. We tested the ability of our model to do so by introducing a change to the Continuity dataset by introducing probes containing an additional pillar (that never was present during training) between the two pillars (that were present during training). Crucially, the physical concept at stake remains the same: continuity of a moving object. We found that the model trained on the Continuity dataset generalized to these new probes (inconsistent probes were significantly more surprising: $(M = 3.47;\ t(7999) = 5.11,\ p < .05)$), although this generalization did not hold for the jointly trained model ($M = 0.17;\ t(7999) = 1.58,\ p = .06$).

Additionally, for the model trained only on Unchangeableness, we tested generalization to a maximum of two more objects than present during training. We found significantly more surprise for inconsistent probes when we modified position ($M = 2.48;\ t(7999) = 136.25,\ p < .05$), color ($M = 0.98;\ t(7999) = 70.44,\ p < .05$), and shape ($M = 1.94;\ t(7999) = 98.60,\ p < .05$).

Our initial choice to employ a large training set was aimed at assuring sufficient diversity in the training data. However, we also examined the sensitivity of VOE effects to sample size. Analyses using the Object Persistence dataset revealed significant VOE effects with as few as ten thousand training examples. As illustrated in Figure 2.5, the effect of increasing training-set size was primarily to reduce the variance of VOE effects across independently sampled test sets.

Next, we examined the results of training the model on all five datasets concurrently, interleaving examples and applying the same evaluation procedure as before. This revealed a statistically significant VOE effect for all categories except Containment (see Figure 2.4). By matching the number of training examples per category, we observed that VOE effects emerged more rapidly for most categories during joint training than separate training (Figure 2.7). Putting all these results together, it appears that concurrent training may give rise to both transfer and interference.

Concurrent with the present work, Riochet and colleagues produced an interestingly related dataset (Riochet et al., 2018). While not framed specifically in terms of VOE, their "IntPhys" dataset is also concerned with the ability of systems to understand intuitive physics and is amenable to our VOE technique. We found that the same model used in our other experiments also showed a VOE effect when trained and tested on their visually quite different "dev" dataset ($M = 0.95;\ t(209) = 2.14,\ p < .05$).

2.7 Discussion

Here we summarize our contributions, related work in the literature, and promising future directions.

2.7.1 Present Work

The initial experiments presented provide a clear demonstration of the potential utility of the VOE method for artifical intelligence, as pioneered in developmental psychology. As we have argued, the appeal of the paradigm lies not only in its focus on surprise, but also in the use of carefully controlled probe stimuli, assessing mastery of specific physical principles. The work presented only scratches the surface of the developmental psychology literature. The latter points to a wide range of phenomena, and provides a rich source of probe designs, which could be leveraged in further work.

Our results are an important step toward developing integrated models for learning intuitive physics, tackling the problem by investigating how far can standard deep learning methods take us in capturing expectations about intuitive physics. In our experiments, we employed a relatively generic model (our only requirements were the ability to compute the KL-divergence and some form of memory) as a baseline for testing. This memory-augmented VRNN displayed a surprising capacity to assimilate basic physical concepts. Whereas it has recently been asserted that "Genuine intuitive physics...is unlikely to emerge from gradient-based learning in a relatively generic neural network" (Lake et al., 2017), the results we have reported suggest it may be important to give this a try, if only to determine where the approach truly breaks down. The present experiments suggest that a fairly generic neural network architecture can deal with significant train-test separation, as well as considerable visual and dynamic diversity, to extract fundamental physical principles. Some potential ways of testing the limits of such successes are noted below.

2.7.2 Future Work

This work presents a small-scale initial application of the VOE method, which provides a foundation for a number of next steps. First, it will be of interest to probe a wider range of concepts, by continuing to draw on the developmental psychology literature related to other concepts such as collisions, number, support, and others (Baillargeon and Carey, 2012a). Second, it will be of interest to further separate our train and test sets, ideally moving to training with more free-form, naturalistic data (Kay et al., 2017), or even video data from infant studies (Clerkin et al., 2017a; Smith and Slone, 2017). As the enterprise expands in this way, we anticipate that more richly structured learning systems may be required, incoprorating such ingredients as attention (Vaswani et al., 2017), relation networks (Santoro et al., 2017a), attentive comparators (Shyam et al., 2017), among others.

We measured surprise as the KL-divergence between the posterior and prior distributions due to its established link to human gaze. However, the possibility for other metrics (e.g. Wasserstein distance) remains.

In order to gain a truly satisfactory knowledge of physics it may also ultimately be necessary to step beyond passive learning to active engagement with the environment, through manipulation and movements that shift the point of view.

Applications of the VOE method in AI work may also hold the potential to feed insights back to developmental psychology. Developmental psychologists are still debating various algorithmic-level explanations (Marr, 1982) underpinning intuitive physics in infants (see the Physical Reasoning System (Baillargeon and Carey, 2012a) for an example). Our framework affords the option to implement and compare these different algorithms and determine which are most successful.

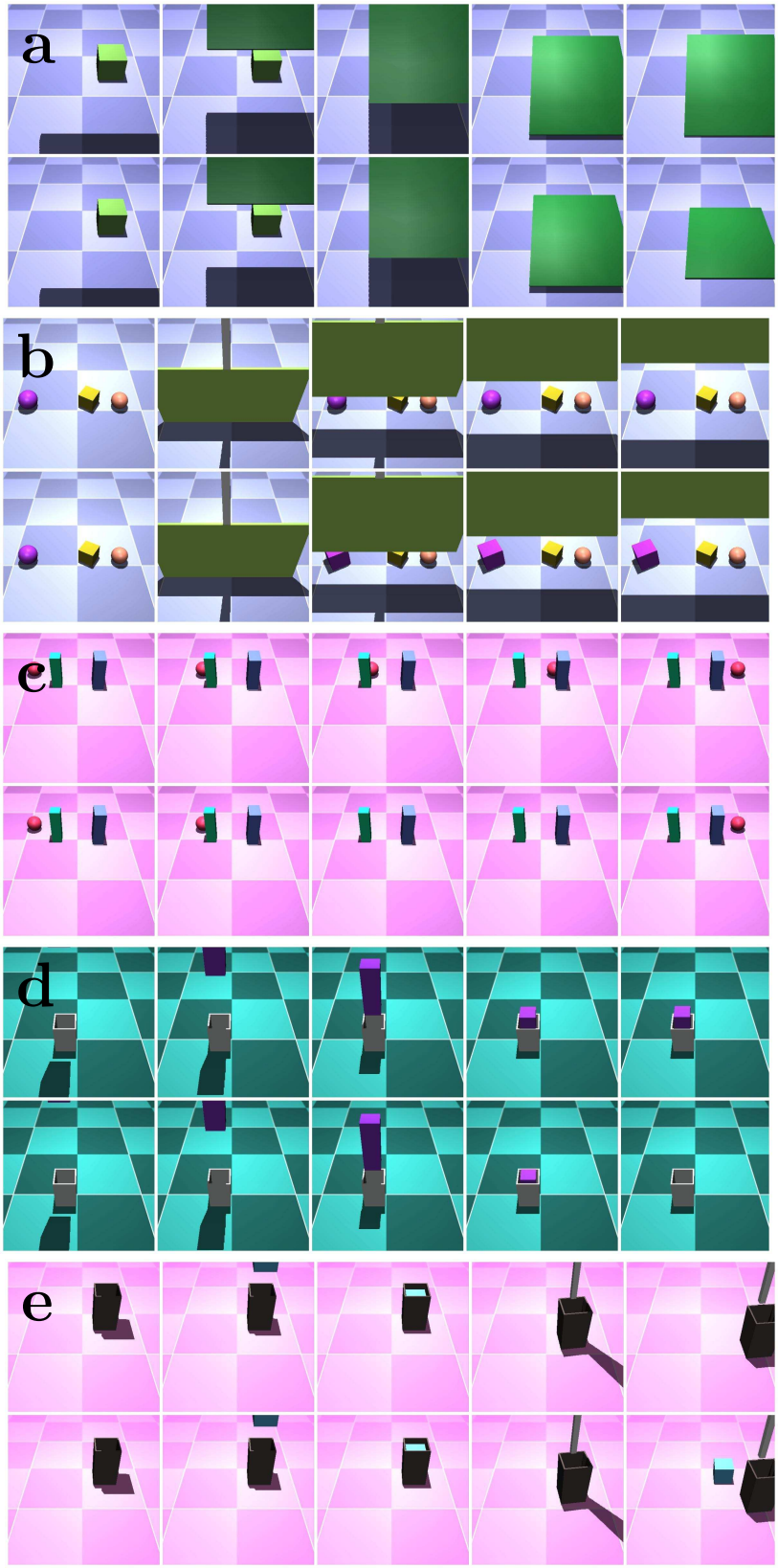


Figure 2.1: Example (subsampled) test probes for each dataset. The first and second rows show consistent and inconsistent probes, respectively. Categories: *a*, Object Persistence; *b*, Unchangeableness; *c*, Continuity; *d*, Solidity; *e*, Containment

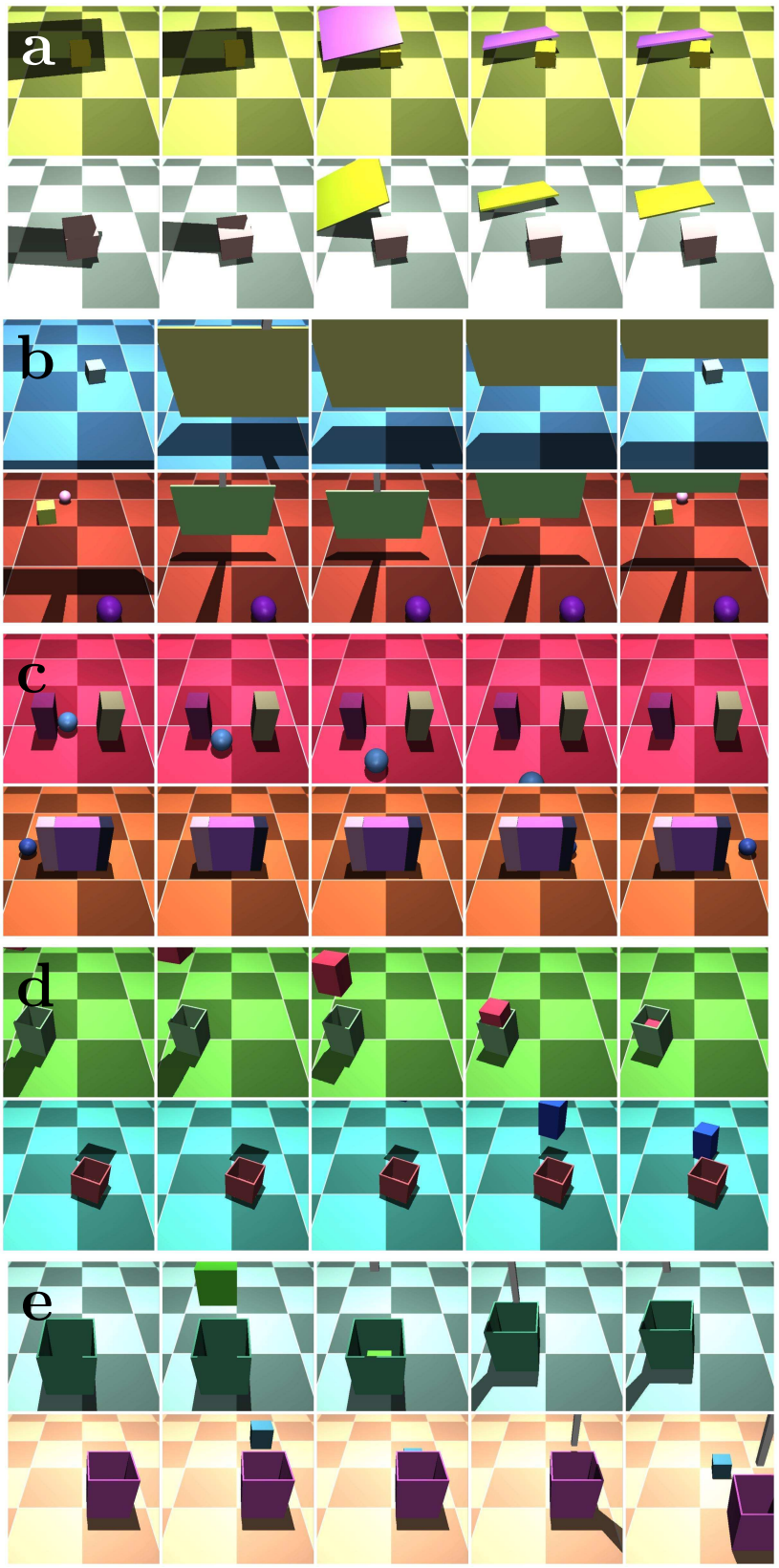


Figure 2.2: Examples of (subsampled) training data (consistent examples and/or controls). Categories: *a*, Object Persistence; *b*, Unchangeableness; *c*, Continuity; *d*, Solidity; *e*, Containment

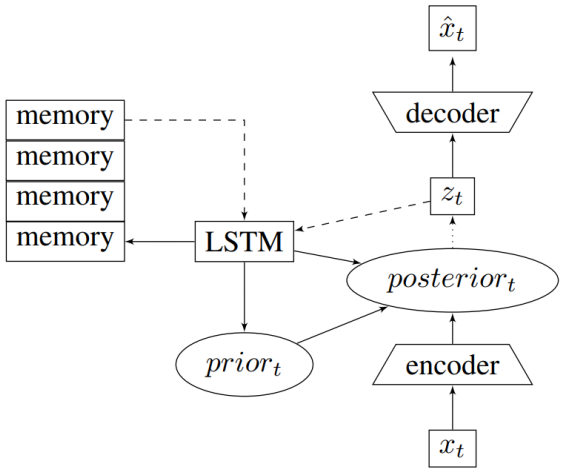


Figure 2.3: Model architecture for a single time step. Dotted lines depict sampling from a distribution and dashed lines denote a one time-step delay. Thus, the sampled z_t and the memory state from the *previous* time step are used to calculate the prior and posterior for the current time step.

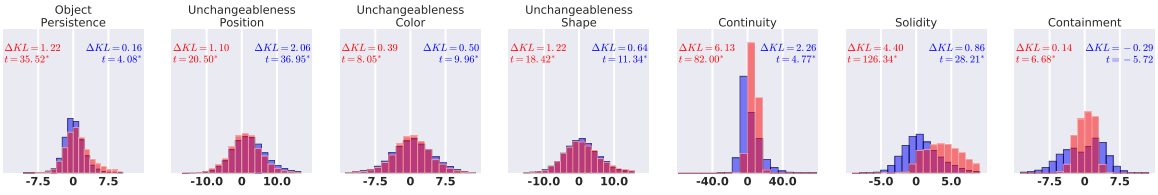


Figure 2.4: Difference in KL divergence (x-axis) over final evaluation set of 8,000 probe pairs. Histograms show the distribution over the evaluation set with outliers >3 standard deviations removed. Text shows the result of significance tests with an asterisk denoting significance. Red corresponds to models trained separately, blue to model trained concurrently.

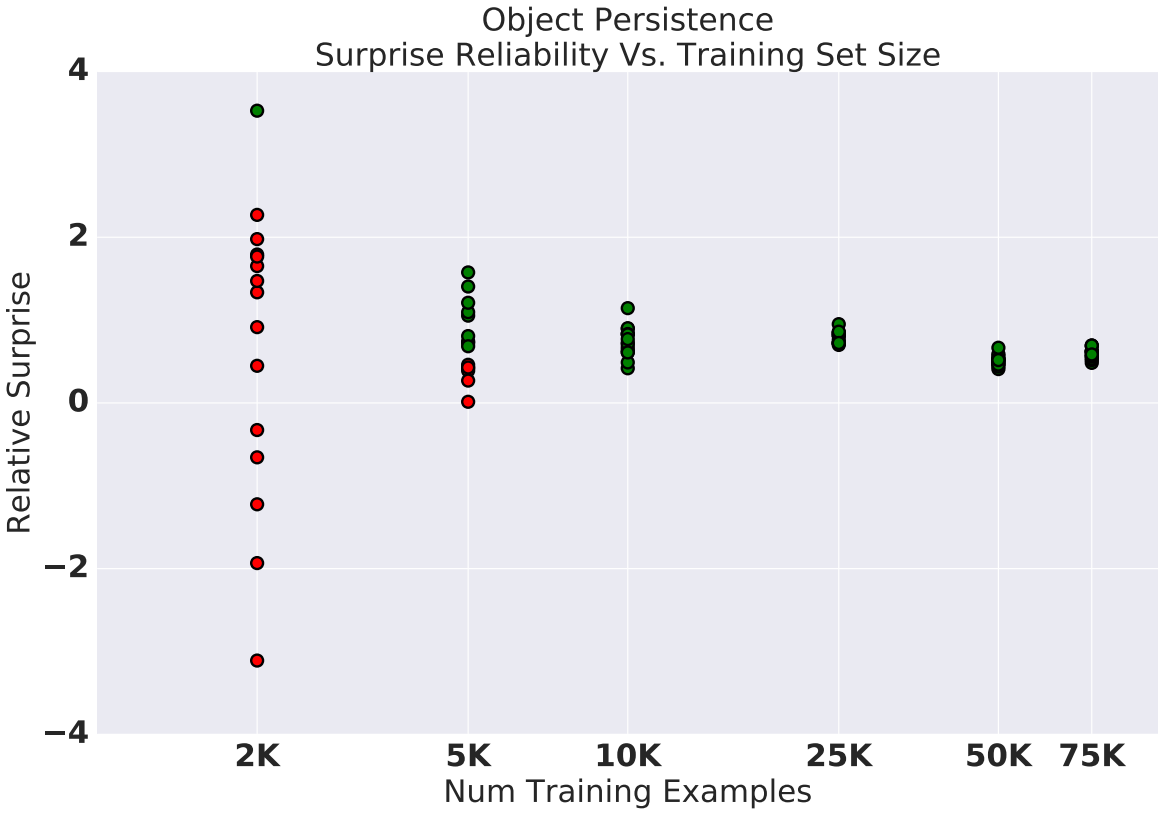


Figure 2.5: Reliability of Relative Surprise (y-axis) for models trained with different Training Set Sizes (x-axis). Each point represents a different evaluation set of 2,000 probe pairs with green points showing a statistically significant effect. As the training set size increases, the variance of the effect size decreases and reliability of statistical significance increases.

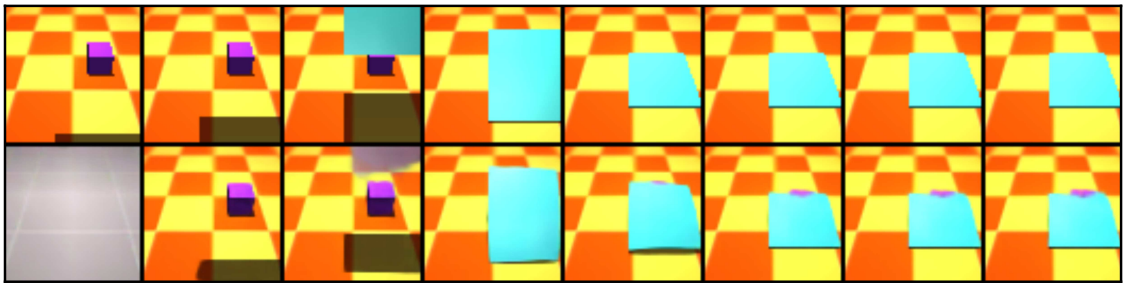


Figure 2.6: Top row: Subsampled frames from an inconsistent probe for the Object Persistence dataset. Here physics is violated when the sheet magically collapses the magenta cube. Bottom row: Network's predictions for the above frame based only on previous frames. Note how it rectifies the input with the correct physics displaying a scene where the sheet rests on top of the magenta cube.

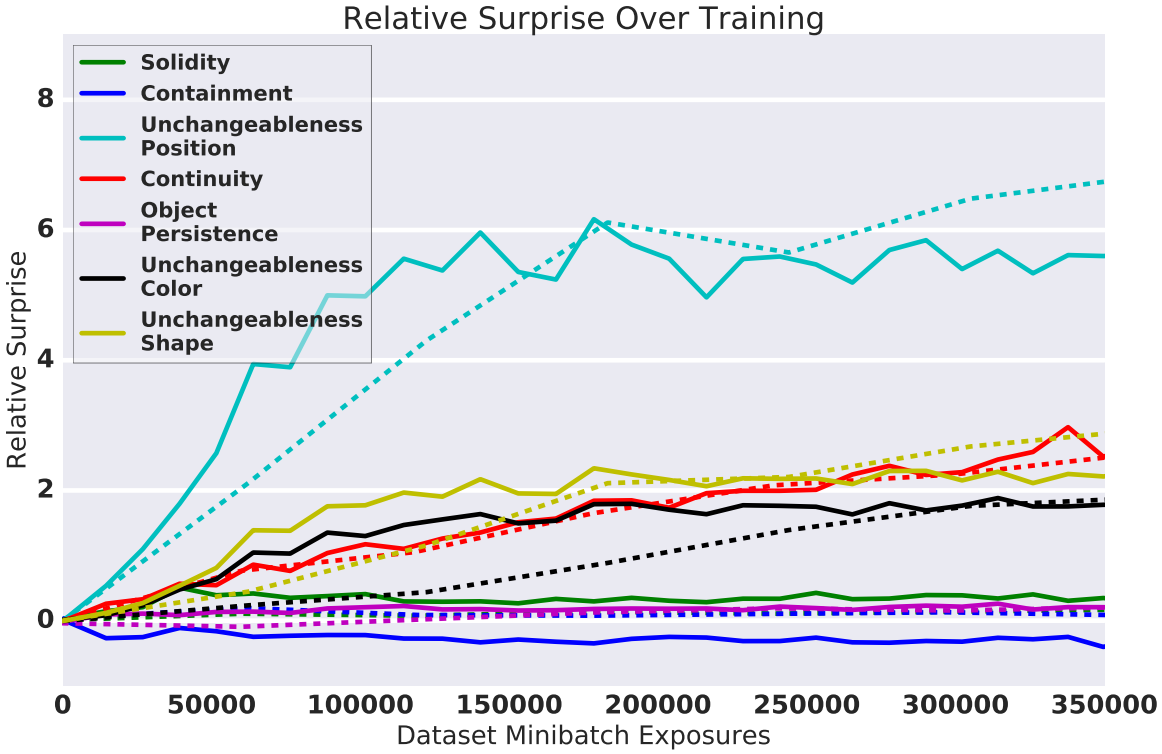


Figure 2.7: Relative Surprise Over Training for Models Trained on Datasets Individually (dashed lines) or Simultaneously (solid lines)

2.8 Appenfix

2.8.1 Convolutional Architecture

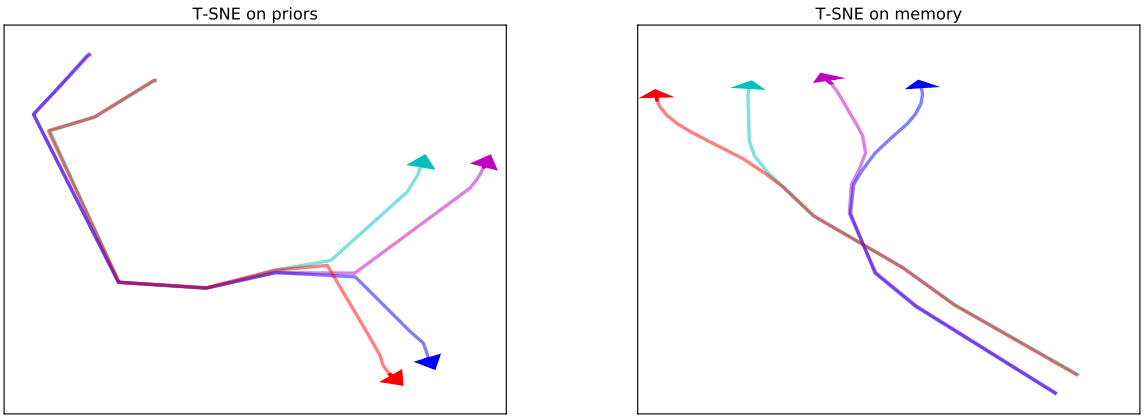
For encoding the images, we used 6 ResNet blocks (He et al., 2016). Each block had three 2D convolutions with kernel sizes = 1x1, 3x3, 1x1. An activation function, rectified linear (ReLU), was only applied after all of the convolutions. The number of channels in the first two convolutions was 64 while the final convolution had 128 output channels. For all convolutions, we used a stride of 2x1. After the series of ResNet blocks, we passed the output through a final MLP with 500 hidden units using a ReLU activation function.

The decoder transposed the encoder architecture to scale from the latent output size to the full observation size. However, instead of doing a deconvolution to increase the dimensionality, we used a size-preserving convolution followed by an upsampling operation.

2.8.2 Model Representations For Consistent Vs. Inconsistent Probes

In Figure 2.8, we show a t-SNE (van der Maaten and Hinton, 2008) representation of the time courses of priors (Figure 2.8a) and memory (Figure 2.8b) for counter-balanced probes in the Object Persistence category. Counter-balanced probes make up 2 pairs of consistent and inconsistent probes, all of which can start and end in only two possible ways. In these plots, the consistent probes are represented by blue and cyan curves, and the inconsistent probes are represented by red and magenta curves. In the prior plot, we can see that initially the red and cyan curves are paired correctly (because they begin identically), as are the blue and magenta. Later in the sequence (moving from the base of the curve in the direction of the arrowhead) this matching is reversed in a manner consistent with the change induced in the inconsistent probes.

Due to the manipulation, the videos represented by red and blue end identically, and likewise for cyan and magenta. When examining memory, we see the same correct initial pairing of colors as with the priors. On the other hand, when the manipulation is introduced we do not see a reversal of the pairing (as we do with the priors), but rather a forking effect. This indicates that the memory maintains information to correctly keep each trajectory distinct based on its entire history, while still able to use the appropriate parts of memory when needed for accurate prior estimation.



(a) Priors (b) Memory

Figure 2.8: t-SNE Embeddings

2.8.3 Model Predictions for Inconsistent Probes

Below we show more examples of the model's predictions for inconsistent probes that include impossible manipulations. Frames are subsampled compared to the actual input received by the model.

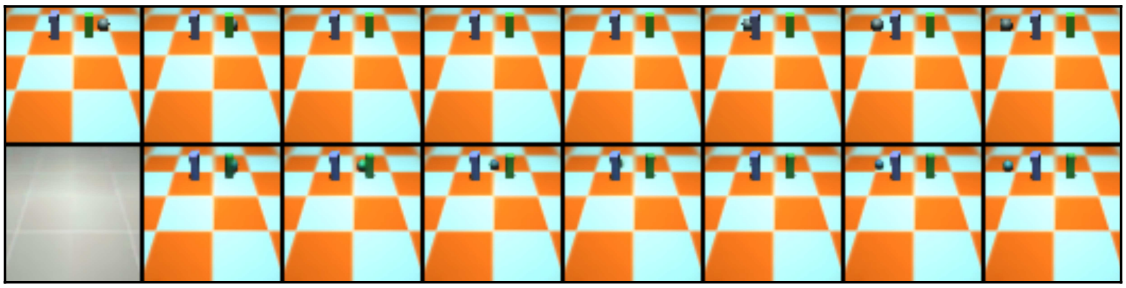


Figure 2.9: Network predictions (bottom row) for Continuity inconsistent probe (top row). Even though the inconsistent probe has the ball disappear between the pillars, the network predicts it should appear.

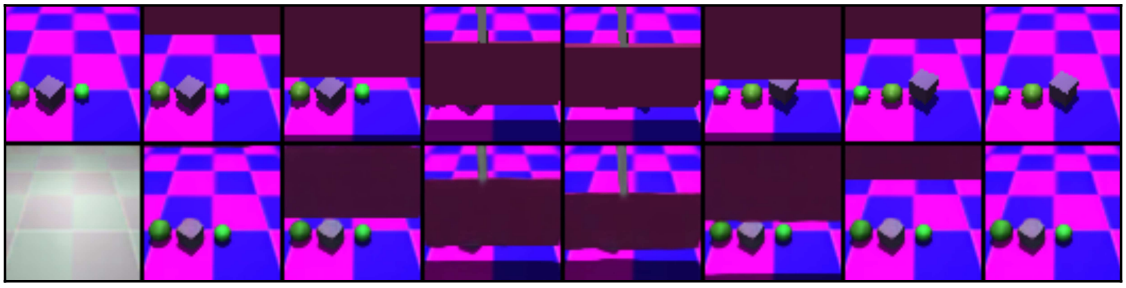


Figure 2.10:) Network predictions (bottom row) for Unchangeableness Position inconsistent probe (top row). Even though the object positions have been swapped after the occluder is lifted, the network predicts the objects should keep their initial placement.

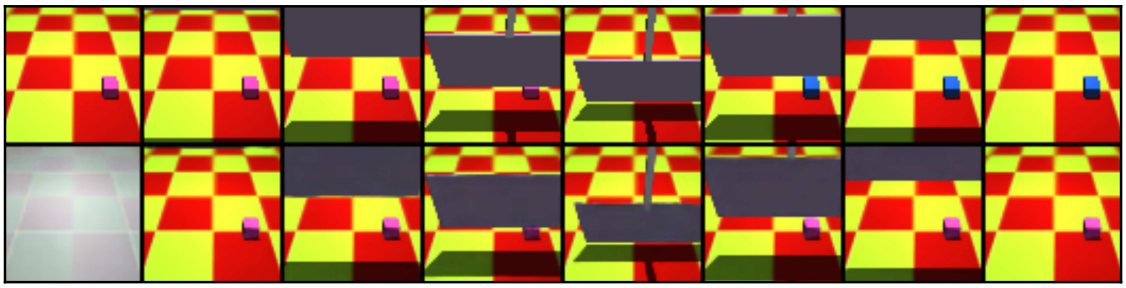


Figure 2.11:) Network predictions (bottom row) for Unchangeableness Color inconsistent probe (top row). Even though the cube has changed color after the occluder is lifted, the network predicts it should keep the original color.



Figure 2.12:) Network predictions (bottom row) for Unchangeableness Shape inconsistent probe (top row). Even though the yellow sphere has been changed to a yellow cube after the occluder is lifted, the network predicts it should remain a sphere.

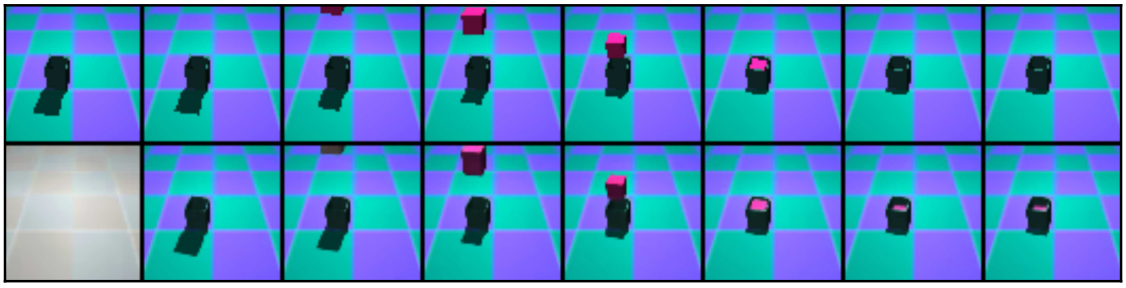


Figure 2.13:) Network predictions (bottom row) for Object Solidity inconsistent probe (top row). Even though the cube falls through the container and floor in the inconsistent probe, the model correctly predicts it should remain in the container.

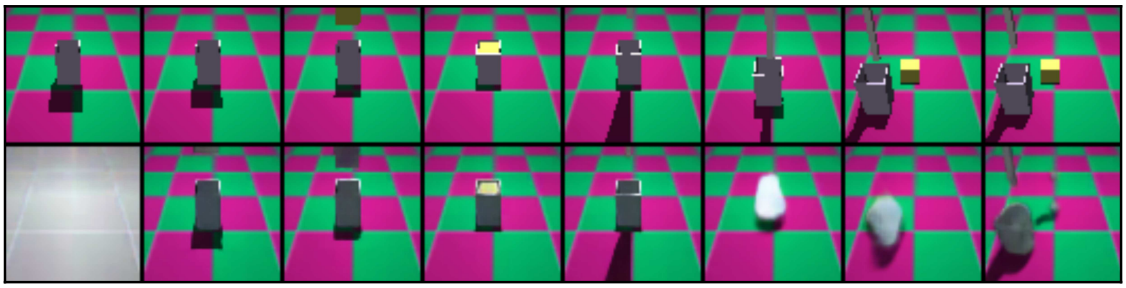


Figure 2.14:) Network predictions (bottom row) for Containment inconsistent probe (top row). When the container is moved, the cube that fell into it magically phases through it, remaining in the place the container used to be. In the first frame after this manipulation, the model correctly predicts nothing should be in the place where the container used to be. However, the model cannot hold onto this correction for an extended period of time and integrates the inconsistent information into the prior, as seen in the final frame where the space that should be empty has a blurred entity in it.

3

Learning Intuitive Physics With Objects

My second project had two goals. First, we sought to improve our virtual VoE paradigm to remove potential biases in the visual statistics of our probe displays. This lead to an improved version of our synthetic dataset where the set of possible and impossible probes were matched both at the level of individual image frames *and* the pairwise transitions between images. This ensured that any observed VoE effects were really due to violations of the principles which govern the dynamics of objects. Second, given the limited success of the unconstrained model from my first project, we aimed to take seriously the nativist claims that object-centric inductive biases enable learning intuitive physics. To that end, we built a deep learning system, PLATO, with object-centric representation and computation. It represented images as objects. It tracked those objects over time. It reasoned about the relations between objects. These inductive biases in our model's architecture instantiated the hypothesized domain-specific constraints on infant learning of intuitive physics. We found that this model was capable of learning intuitive physics in a general way. This model was capable of rich generalization over physical events with just 28 hours of

visual experience. This stood in stark contrast to the narrow generalization observed in our previous project. We ablated the model and found the object inductive biases were critical for learning intuitive physics in our dataset. This project was the first to show a model capable of learning generalizable physical intuition from perceptual data (but assuming access to objects). It demonstrated that within a specific class of systems (our deep learning models) and using a specific formulation of the intuitive physics task (our dataset), access to objects makes the otherwise intractable task of learning intuitive physics tractable. This project is currently being submitted and this chapter is a reproduction of that submission. The working citation for the paper is:

Luis Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Learning intuitive physics: A deep-learning model inspired by developmental psychology. *submitted*, 2021

Learning Intuitive Physics: A deep-learning model inspired by developmental psychology

3.1 Introduction

The field of AI has made astonishing progress in recent years, mastering an increasing range of tasks that now include Atari video games (Mnih et al., 2013), board games such as chess and Go (Silver et al., 2017), and scientific problems including protein-folding (Senior et al., 2020). At the same time, success in these narrow domains has made it increasingly clear that something fundamental is still missing. In particular, state-of-the-art AI systems still struggle to capture the 'common sense' knowledge that guides prediction, inference and action in everyday human scenarios (Lake et al., 2017; Smith, 2019). In the present work, we focus on one particular domain of common sense knowledge: *intuitive physics*, the network of concepts that underlies reasoning about the properties and interactions of macroscopic objects (Michotte, 1963). Intuitive physics is fundamental to embodied intelligence, most obviously because it is essential to all practical action, but also because it provides one foundation for conceptual knowledge and compositional representation in general (Johnson, 2013). Despite considerable effort, however, the recent advances in AI have yet to yield a system that displays an understanding of intuitive physics comparable to that of even very young children.

To pursue richer common sense physical intuition in AI systems, we take, at multiple points in our work, inspiration from developmental psychology, where the acquisition of intuitive physics knowledge has been an intensive focus of study (Baillargeon and Carey, 2012b; Baillargeon, 2008; Spelke et al., 1992; Spelke and Kinzler, 2007). We build a deep-learning system that integrates a central insight of the developmental literature, which is that physics is understood at the level of discrete *objects*

and their interactions. We also draw on developmental psychology in a second way, which relates to the problem of *behaviorally probing* whether an AI system (or in the case of developmental psychology, an infant or child) possesses knowledge of intuitive physics.

In developing behavioral probes for research on children, developmental psychologists have based their approach on two principles. First, that the core of intuitive physics rests upon a set of discrete concepts (Baillargeon et al., 1985a; Spelke et al., 1992) (e.g. object persistence, object solidity, and continuity, etc.) which can be differentiated, operationalized and individually probed. By specifically targeting concept acquisition, our work is quite different from standard approaches in AI for learning intuitive physics which measure progress via video or state prediction (Watters et al., 2017; Agrawal et al., 2016; Chang et al., 2017b) metrics, binary outcome prediction (Lerer et al., 2016), question-answering performance (Yi* et al., 2020; Girdhar and Ramanan, 2020) or high reward in reinforcement learning tasks (Veerapaneni et al., 2020). These alternative approaches intuitively seem to require an understanding of some aspects of intuitive physics, but do not clearly operationalize or strategically probe an explicit set of such concepts.

The second principle used by developmental psychologists for probing physical concepts is that possession of a physical concept corresponds to forming a set of expectations about how the future can unfold. If human viewers expect that objects will not "wink out of existence" when they are out of sight, then they have the concept of *object persistence* (Baillargeon, 1987). If they expect that objects will not interpenetrate one another, then they have the concept of *solidity* (Hespos and Baillargeon, 2001c). If they expect that objects will not magically teleport from one place to another, but instead trace continuous paths through time and space, then they have the concept of *continuity* (Spelke et al., 1992). With this conceptual

scaffolding, a method for measuring knowledge of a specific physical concept emerges: the violation-of-expectation (VoE) paradigm (Baillargeon, 1987).

Using the VoE paradigm to probe for a specific concept, researchers show infants visually similar arrays (called *probes*) that are either consistent (*physically possible*) or inconsistent (*physically impossible*) with that physical concept. If infants are more surprised by the possible array (as typically assessed via gaze duration (Baillargeon, 1987)), this provides evidence their expectations, derived from their knowledge of the probed physical concept, were violated. As an example, consider the concept of continuity (depicted in Figure 3.1): objects trace a continuous path through time and space. For the possible probe (Figure 3.1, first row), researchers (Aguiar and Baillargeon, 1999) showed an object moving horizontally behind a pillar, being occluded by that pillar, subsequently emerging from occlusion, and traveling towards a second pillar, where it was again occluded behind that pillar and emerged from occlusion one final time. In the impossible probe (Figure 3.1, third row), when the object is occluded by the first pillar, it does not emerge from occlusion immediately. Instead, after some delay, the object emerges from behind the second pillar - never appearing in the space between the two pillars and thus seeming to teleport from one pillar to the other. Experiments with infants have shown that by the age of 2.5 months, they gaze longer at an object that teleports between two screens than an object that moves continuously from one screen to the next (Aguiar and Baillargeon, 1999). This same strategy has been used by developmental researchers to accumulate strong evidence that infants acquire a wide range of distinct physical concepts (Baillargeon and Carey, 2012b; Baillargeon, 2008; Spelke et al., 1992; Spelke and Kinzler, 2007) within the first year of life.

In previous work (Piloto et al., 2018), we introduced a machine-learning video dataset designed to systematically test how well models can learn specific physical concepts (see Discussion section for comparison against similar datasets developed in

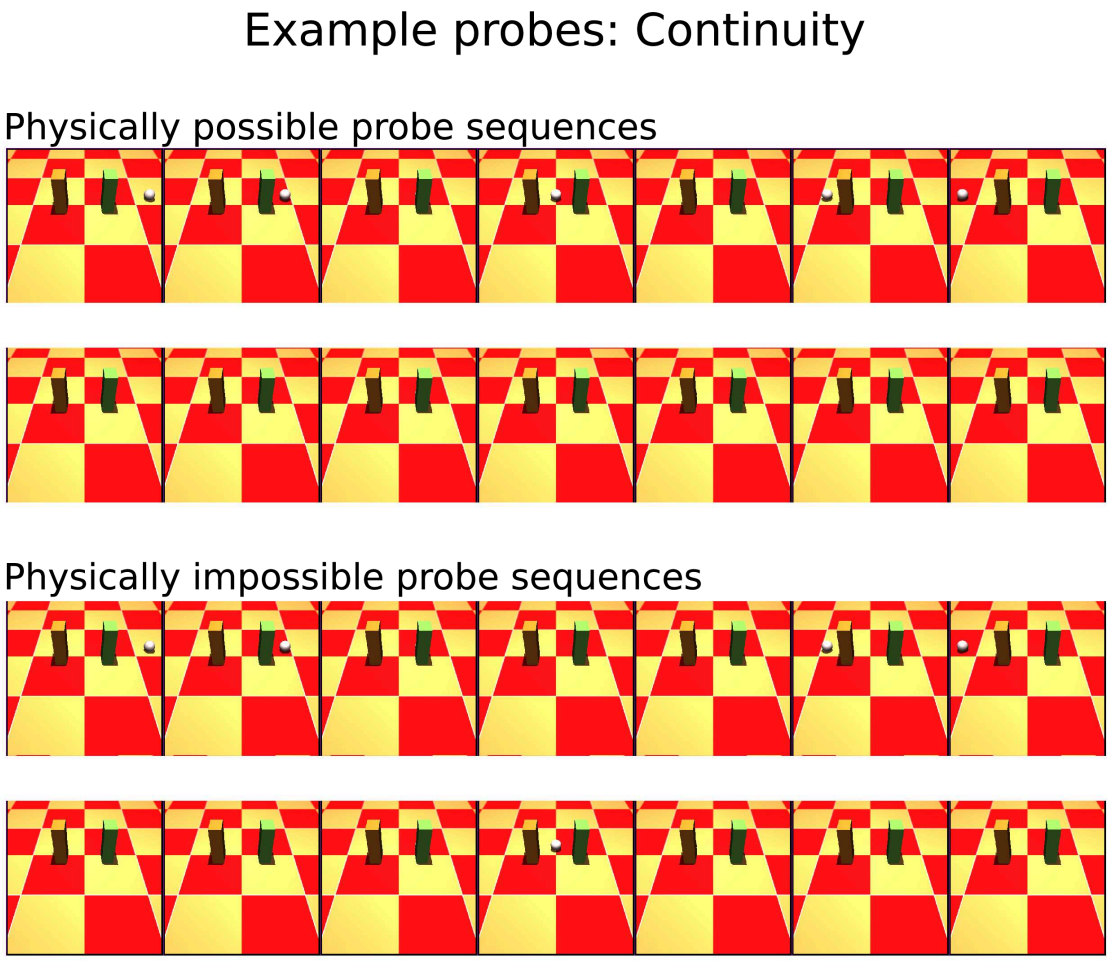


Figure 3.1: Example probes adapted from (Aguiar and Baillargeon, 1999) to assess the physical concept of *continuity* (Spelke et al., 1992; Baillargeon, 2002): objects trace a continuous path through time and space. Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), a ball rolls behind two occluders. In the second possible probe (second row), no ball is present. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the ball rolls behind the first occluder and emerges from the second occluder, never appearing between the two occluders. The second impossible probe (fourth row) has the opposite structure: the ball appears between the two occluders, but was never seen rolling behind the first occluder or rolling out of the second occluder.

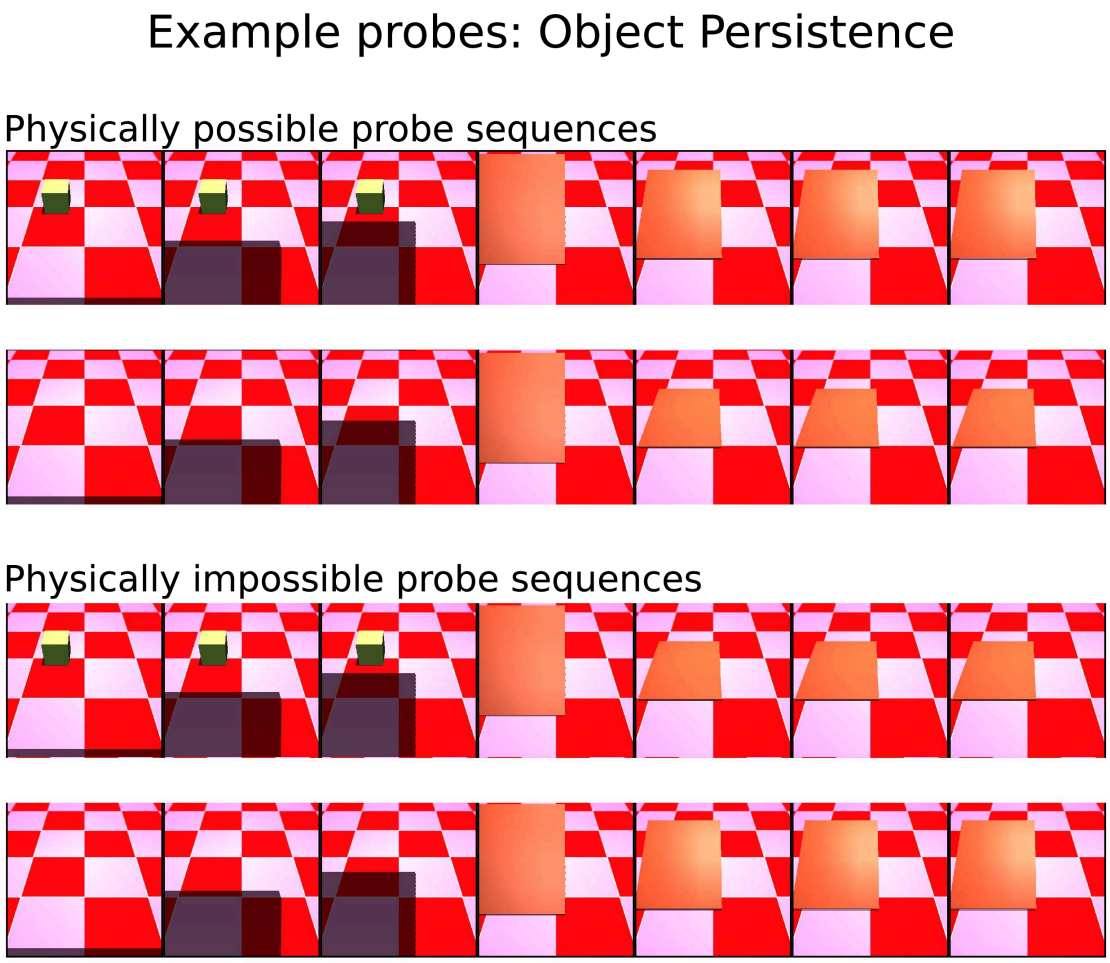


Figure 3.2: Example probes adapted from (Baillargeon, 1987) to assess the physical concept of *object persistence* (Baillargeon, 2008). Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), a rigid plank falls onto, occludes, and is propped up by a cube. In the second possible probe (second row), the plank falls and remains flush against the ground. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the plank falls onto the cube. However, instead of being propped up by the cube as you would expect, the plank ends up flush against the ground as if no cube existed at all. The second impossible probe (fourth row) shows a plank falling where there is no cube. Although it should end up flush against the ground, the plank comes to rest at an angle as if it landed on a cube.

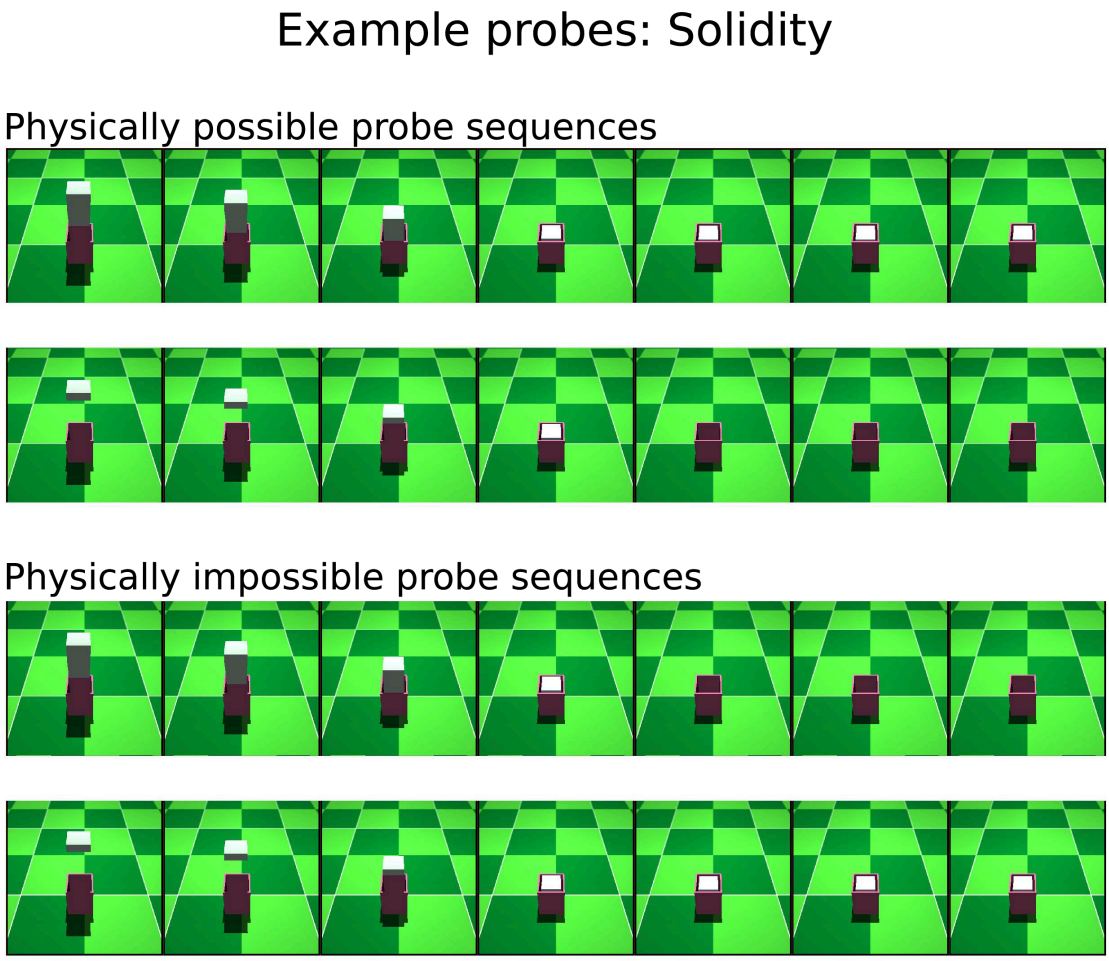


Figure 3.3: Example probes adapted to assess the physical concept of *solidity* (Hespos and Baillargeon, 2001c): objects do not interpenetrate. Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), a tall block falls into a visibly empty container. In the second possible probe (second row), a short block falls into the container. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the tall block continues to fall into the container as if there was no bottom. In the second impossible probe (fourth row), the short block falls into the visibly empty container but fails to fall all the way in.

parallel (Riochet et al., 2018) and subsequently (Smith et al., 2019)). Our original dataset contained procedurally-generated VoE probes that assess acquisition of a set of physical concepts, constructed so that no individual video frame could explain any resulting VoE effect. In the present work, we introduce a much richer video corpus, the *Physical Concepts* (PC) dataset (URL for dataset will be filled in here). This new dataset contains VoE probe videos targeting five physical concepts that have been identified as central in the developmental psychology literature. The first three, continuity (Figure 3.1), object persistence (Figure 3.2), and solidity (Figure 3.3), were introduced above. The fourth concept, 'unchangeableness' ((Káldy and Sigala, 2004; Baillargeon and Carey, 2012b), Figure 3.4), captures the notion that certain object properties (e.g. shape) do not change. The fifth and final concept, directional inertia ((Spelke, 1994), Figure 3.5), involves the expectation that moving objects undergo changes in direction consistent with the principles of inertia. As detailed in the Methods section, each video showing a violation of these physical principles is matched with a corresponding video that provides a physics-consistent baseline, maintaining precisely the same single-image statistics across videos.

Critically, the Physical Concepts dataset also includes a separate corpus of videos intended as *training* data. These videos show a wide variety of procedurally generated physical events (see Figure 3.6 and Methods) involving objects similar to those involved in the probe videos, among others, but never showing the specific events involved in the test probes. Readers are encouraged to view examples in video format: tiny.cc/phys_concepts1 and tiny.cc/phys_concepts2.

Equipped with this dataset and evaluation framework, we now turn to the primary objective of the present research: to build a model capable of learning intuitive physics. Our architecture is inspired by accounts from developmental psychology which posit that three object-centric processes underpin infant intuitive physics behavior (Baillargeon and Carey, 2012b). We leverage recent advances in AI to instanti-

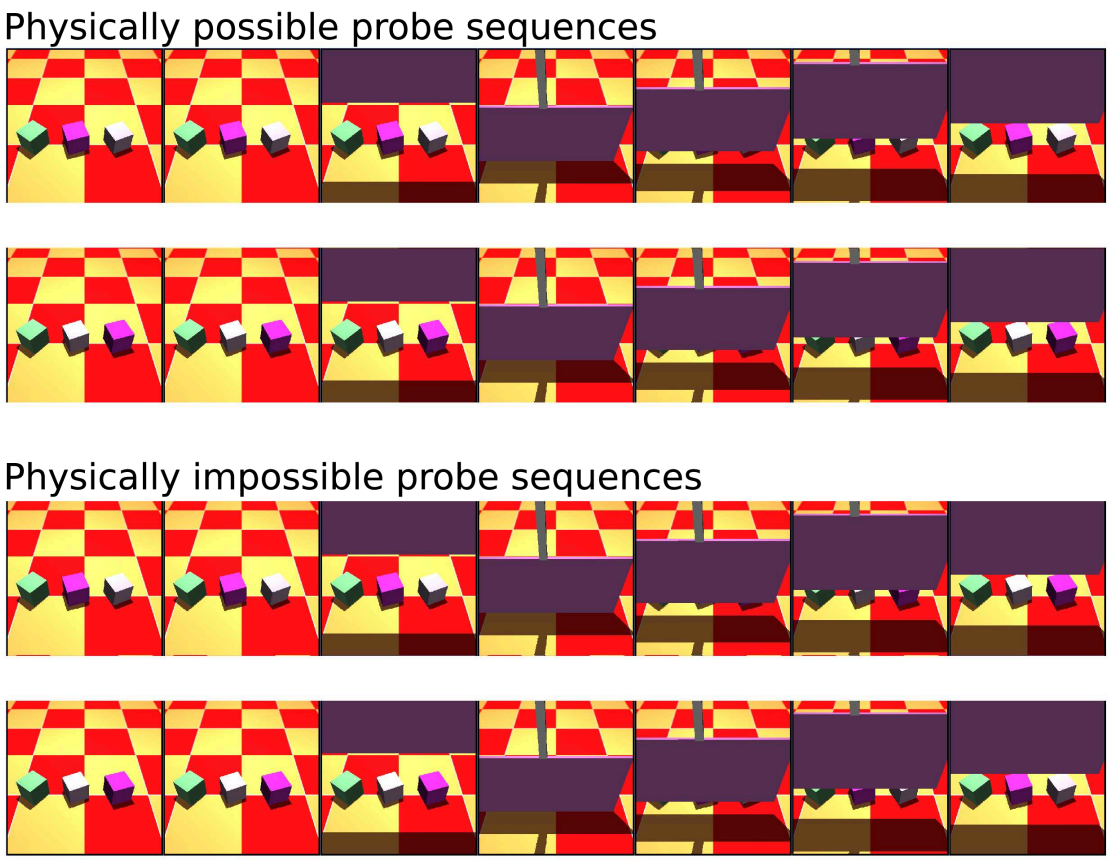


Figure 3.4: Example probes adapted from (Káldy and Sigala, 2004) to assess the physical concept of *unchangeableness* (Baillargeon and Carey, 2012b; Spelke et al., 1992): objects have certain immutable attributes (e.g. they do not change shape, size, or color). Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), an arrangement of objects is occluded by a descending curtain and eventually unoccluded to reveal the initial configuration of objects. In the second possible probe (second row), the same exact objects appear in a different configuration. They are then occluded and unoccluded as in the first probe. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the objects begin in the configuration of the first probe. However, when the occluder is lifted, the objects are seen in the configuration of the second probe, suggesting that the objects have somehow changed shape, size, or color while occluded. The second impossible probe (fourth row) starts with the configuration from the second possible probe and ends with the configuration from the first possible probe.

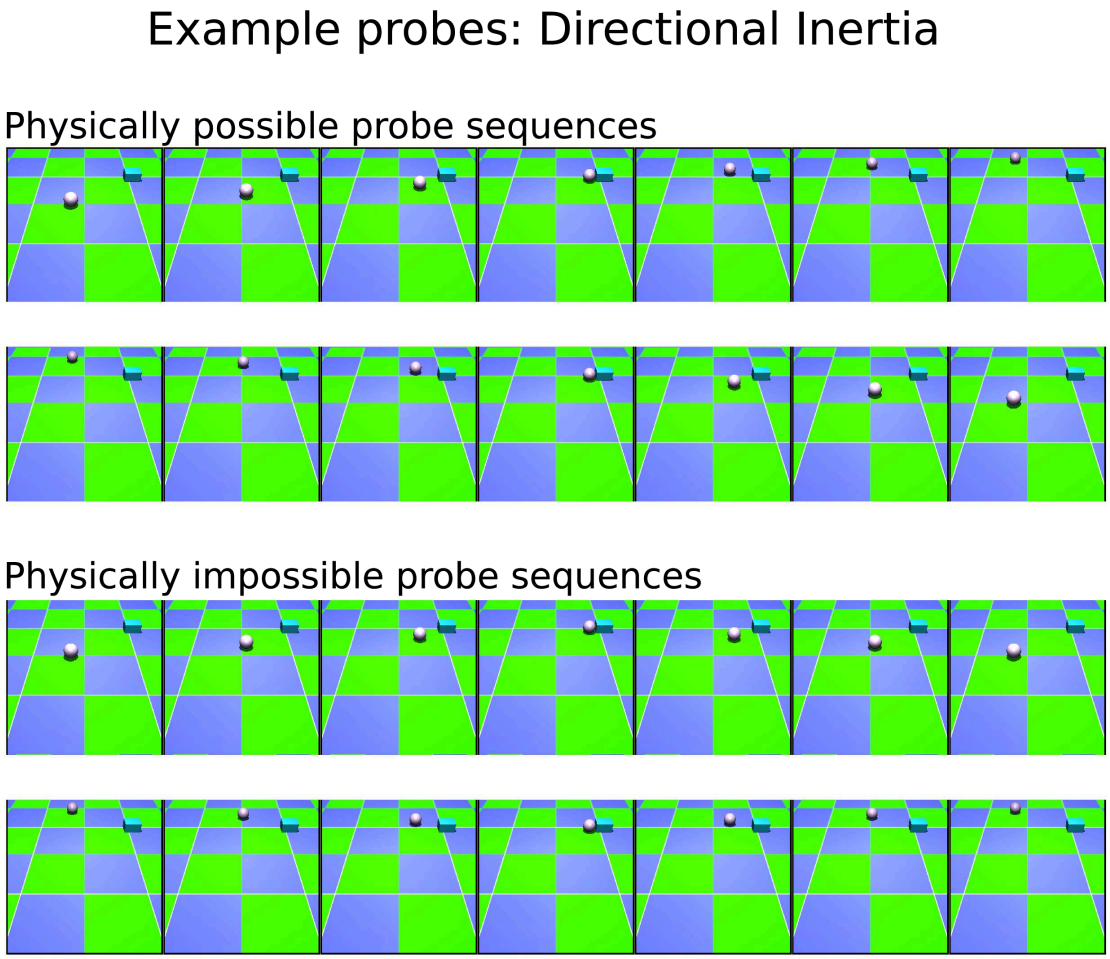


Figure 3.5: Example probes adapted from (Spelke et al., 1994a) to assess the physical concept of *directional inertia*: colliding objects undergo changes to the direction of their velocity in accordance with the angle of incidence. Each row corresponds to one temporally down-sampled video in a probe tuple. Actual videos consist of a total of 15 frames. The top two rows are physically possible probes and the bottom two rows are physically impossible probes. **Physically Possible Probes**: In the first physically possible probe (first row), a ball rolls towards a block, ricochets off the block (which reverses its horizontal velocity) and maintains its movement towards the rear of the scene. The second possible probe (second row), reverses this trajectory: it starts at the rear of the scene, ricochets horizontally off the block, and continues to roll towards the front of the scene. **Physically Impossible Probes**: These probes are formed by splicing parts of the physically possible probes into impossible events. In the first impossible probe (third row), the ball rolls from the front of the scene towards the rear. When it collides with the block, instead of continuing towards the back of the scene, it reverses its direction back to the front of the scene. Similarly, in the second impossible probe (fourth row), when the ball collides with the block, it unexpectedly rolls back in the direction it came from.

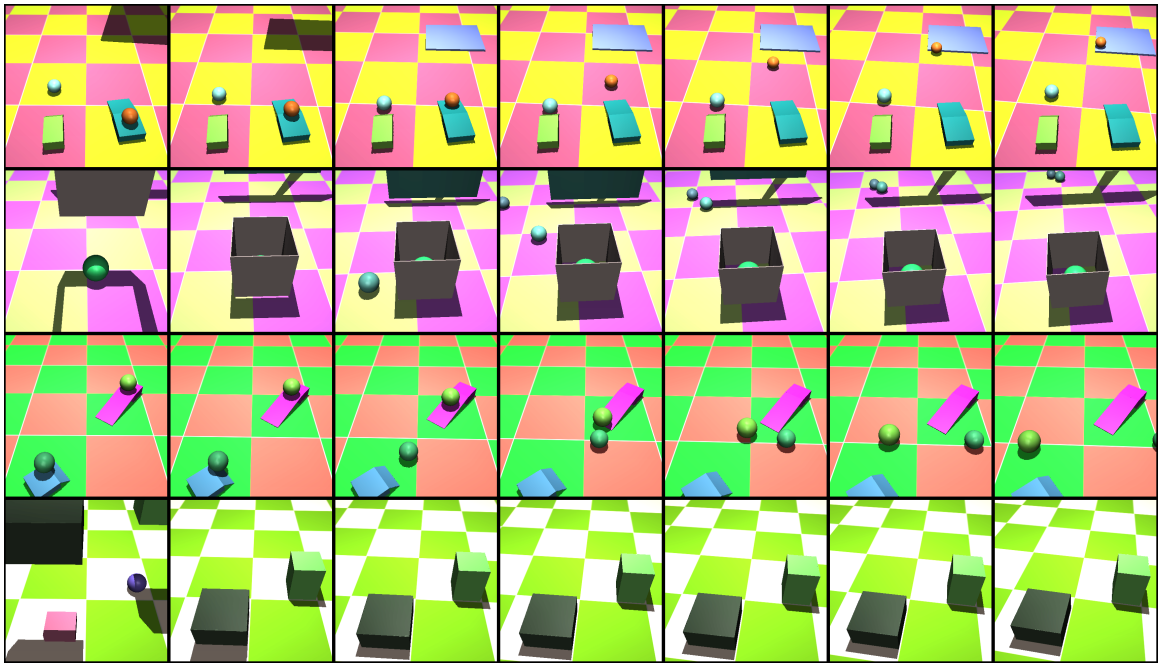


Figure 3.6: Example training videos (temporally downsampled, actual videos contain 15 frames). Scenes are constructed procedurally with composable interactions: objects added to a scene are either initialized completely randomly or target preexisting objects in the scene. Camera starts from a fixed location and drifts randomly over time.

ate these systems in a model, Physics Learning through Auto-encoding and Tracking Objects (PLATO). First and foremost is the process of *object individuation* (Spelke et al., 1992). Object individuation carves the continuous perceptual input of vision into a discrete set of entities, where each entity has a corresponding set of attributes. In PLATO, each segmented video frame is decomposed into a set of object codes via a perception module (Figure 3.7, Left), thus implementing a mapping from visual input to individuated objects. Second, *object tracking* (or *object indexing*) assigns an index to each object enabling a correspondence between object percepts across time (Kibbe and Leslie, 2013; Spelke et al., 1995; Wynn, 1992). In PLATO, the object codes are accumulated and tracked over frames in an object buffer (Figure 3.7, Right), thus implementing object tracking. The final component is *relational processing* of these tracked objects. According to the "Physical Reasoning System" of (Baillargeon and Carey, 2012b), initial object representations are dynamically processed yielding new

representations which are inflected by their relationship to and interactions with other objects. Behavioral evidence that infants compute object relations comes from developmental trajectories showing infants first attend to relative depth to make occlusion judgments, but gradually learn other relevant relations such as relative height (Baillargeon and Carey, 2012b). In PLATO, we learn interactions (3.7, Right) between the object buffer (an object-based long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997a)), and the current percept to produce per-object predictions for the next video frame and update the object-based memory. We train PLATO on a next-step prediction task and evaluate its performance on our suite of intuitive physics probes.

To foreshadow our results, we find that PLATO displays strong VoE effects across all five concept-probe categories in our dataset. By contrast, carefully controlled comparison models that lack object-centered representations fail to achieve above-chance results on the evaluation suite, even when furnished with more computational capacity. Finally, we report that our object-centered model can develop robust VoE effects with a surprisingly small amount of training data, equivalent to 28 hours of visual experience.

3.2 Results

3.2.1 Models

To implement the object-centric approach, our model comprises two main components: a feedforward perceptual module (Figure 3.7, left) and a recurrent dynamics predictor with per-object memory (Figure 3.7, right). The perceptual module takes as input an image and segmentation mask – we discuss later how segmentations can also be learned from scratch from visual data – and converts these into a vector embedding using standard deep-learning autoencoding methods (see Figure 3.7, Left;

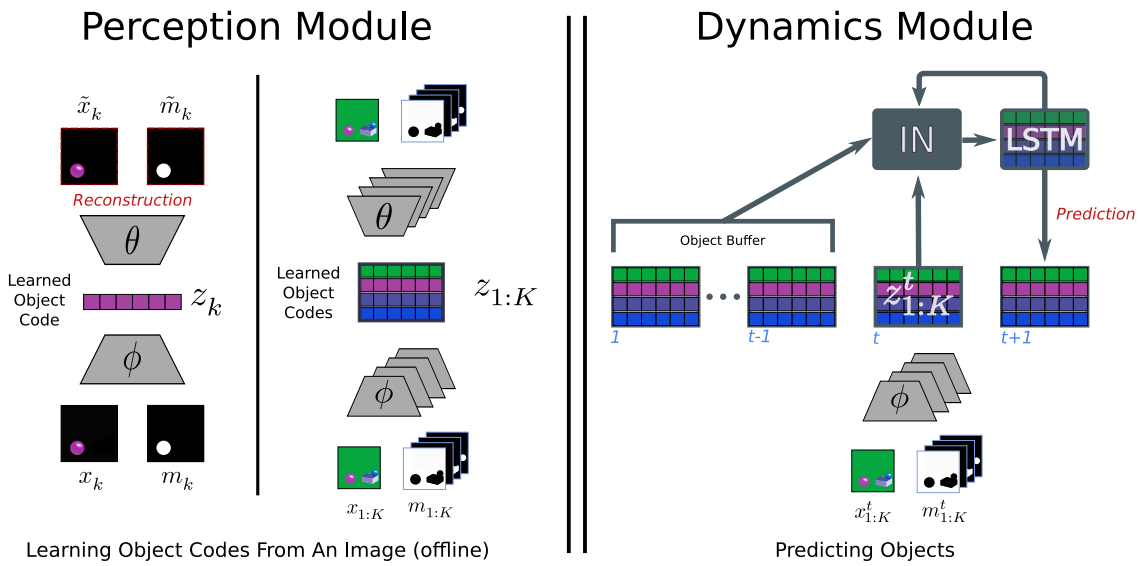


Figure 3.7: Model architecture consists of two components: the perception module (left) and the dynamics predictor (right). The perception module takes a segmented image $(x^t, m^t_{1:K})$ and is pre-trained to reconstruct each object individually (left) using a variational autoencoder with shared parameters: Φ, θ. In doing so, Φ can produce a set of object codes, $z^t_{1:K}$, which represent static object properties in a learned representational space. For training the dhynamics model, the object codes are aggregated over the course of a video into an object buffer. An interaction network (IN) then computes interactions from an object-based LSTM to the object buffer and the current frame's object codes. The result of these interactions are used to update the LSTM and make per-object predictions.

see Methods). In effect, the perceptual module parses the high-dimensional visual input into a small set of discrete object codes.

The dynamics predictor centers on a structured recurrent neural network we developed called an InteractionLSTM (Figure 3.7, Right: 'IN' and 'LSTM' boxes; see Methods). This takes as input the history of single-frame object-level embeddings (Figure 3.7, 'object buffer'), and predicts the set of object codes at the next time step (Figure 3.7, 'object predictions').

We hypothesized, based on the insights provided by developmental psychology, that the involvement of object-level coding would be critical to the acquisition of intuitive physics concepts and corresponding VoE effects. To evaluate this claim, we constructed a well-matched object-agnostic "flat" model as a baseline for comparison.

For this, we took each component in PLATO that used a set of object codes, and replaced that set with a single vector embedding for the entire scene (see Methods). We investigate two versions of the resulting flat model on the basis that we can match either the number of tunable parameters in PLATO or the number of units representing each video frame. In the *Flat: equal parameters* (FEP) baseline, the flat embedding has the same number of dimensions (16) as one of PLATO's single-object codes. This model has the same number of learnable parameters in the dynamics predictor as PLATO, but reduced capacity for representing images in the perceptual module. In the second version, *Flat: equal capacity* (FEC), we give the flat code the same capacity for representing images in the perceptual model (a single 128-dimensional vector, given that PLATO contains eight 16-dimensional object-code slots). In this case the dynamics predictor contains approximately four million more parameters than PLATO (see Methods for full implementational details).

3.2.2 Training and Testing

For all models, training consisted of two phases. In the first phase (Figure 3.7, Left and Middle columns), the perceptual module was trained to reconstruct individual images from the training data (see Methods for loss function and other details). After this phase, the weights within the perceptual module were fixed, yielding an encoder that could produce object (or flat) codes from inputs and a decoder which could take object (or flat) codes and produce an image. In the second phase (Figure 3.7, Right column), we trained the dynamics module to predict the next set of object codes in videos from our training set (using teacher-forcing as detailed in the Methods section).

We evaluated our models using the VoE paradigm for the five physical concepts targeted in the PC test set: Continuity, Directional Inertia, Object Persistence, Solidity, and 'Unchangeableness.' (see Methods section for description). For a single physical concept, we procedurally generated 5,000 probe tuples - each comprising

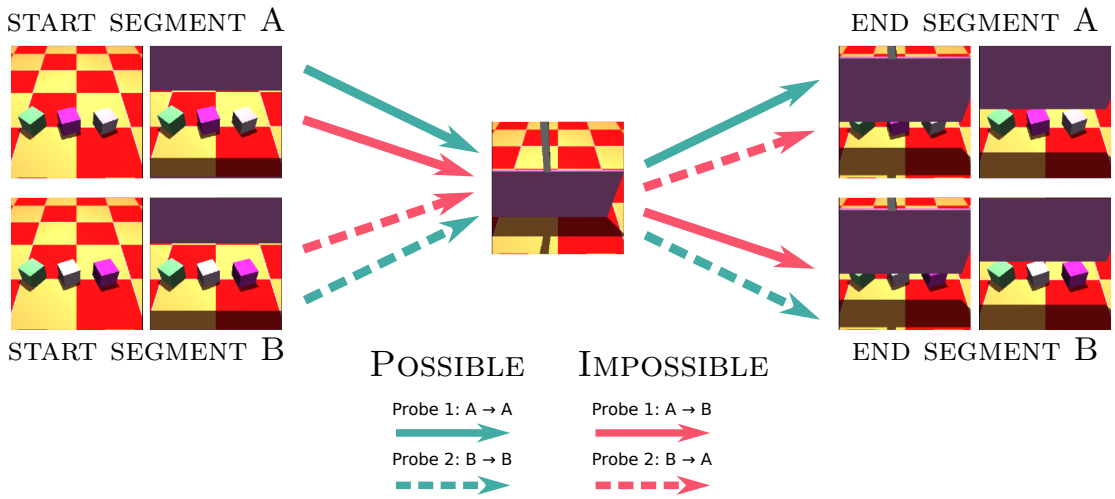


Figure 3.8: Illustration of the splicing procedure used to create a probe tuple for the physical concept of 'unchangeableness' as seen in Figure 3.4. Probe tuples consist of two possible videos and two impossible videos. The possible videos are created by running a physics engine with related initial conditions (e.g. same objects but different positions). Importantly, videos are constructed so they share a frame in common (this is the center image in the figure above). The segments before this shared frame are called the "start segment" (left part of image) and the ones after are called the "end segment" (right part of image). Thus to create any probe, we can gather frames from a start segment, add the common frame, and add frames from an end segment. Possible probes (green arrows above) are formed by starting and ending at corresponding start and end segments (e.g. start segment A to end segment A). Impossible probes (red arrows above) are formed by using a mismatched end segment (e.g start segment A to end segment B). This design ensures that the possible videos and impossible videos within a probe tuple are perfectly matched in terms of individual frames and pairs of frames.

two physically possible probes and two physically impossible probes. The impossible probes were constructed by splicing (illustrated in Figure 3.8) together frames from the possible probes in a way that clearly violates physics. This splicing procedure ensures that the same set of images is present in the possible probes as in the impossible probes. The only difference is the ordering (which yields aphysical events in the impossible probes), an approach pioneered by Riochet et al. (Riochet et al., 2018). Furthermore, we splice frames such that all adjacent frames are physically possible, even if the scene as a whole is not. This ensures that the exact same *pairs* of images appear in both probe types. This precludes a model from showing strong VoE

effects merely based on temporally local inconsistencies (see Supplement for empirical validation).

Under the VoE paradigm, physical concept acquisition is quantified by comparing surprise on the two probe types. To calculate surprise for a given video, we compute for each frame the model's prediction error, defined as the sum-squared error of the system's pixel level prediction. Then, we sum prediction errors across all frames within a video. For each of the 5,000 probe tuples for a physical concept we compute the sum of the surprises on the possible probes (called the *physically-possible surprise*) and similarly compute the *physically-impossible surprise*. We compute an accuracy score where a probe is "classified" correctly if the impossible surprise is greater than the possible surprise. We use the average accuracy to assess the model's acquisition of a physical concept. Whereas accuracy is binary, we can also compute the *relative surprise*, the difference between the impossible surprise and possible surprise, to quantify the magnitude of the surprise effect. To allow for comparison across the probe tuples, we normalize the relative surprise by the sum of both the possible and impossible surprises. This normalization accounts for the fact that some initial conditions yield higher baseline surprises across both probe types (e.g. probes with higher velocities). Finally, to accommodate variability in simulation results, we computed average accuracy and average relative surprise for five different seeds of each model.

3.2.3 Test set performance

At test time, PLATO displayed robust VoE effects in all five probe categories. This was evident both in effect size (relative surprise) which is most analogous to the original developmental experiments (Figure 3.9, top row) and classification accuracy (Figure 3.9, middle row). Time-courses of relative surprise indicated that this rose at appropriate moments in each probe video, coinciding with the onset of the relevant physically impossible event (Figure 3.9, bottom row).

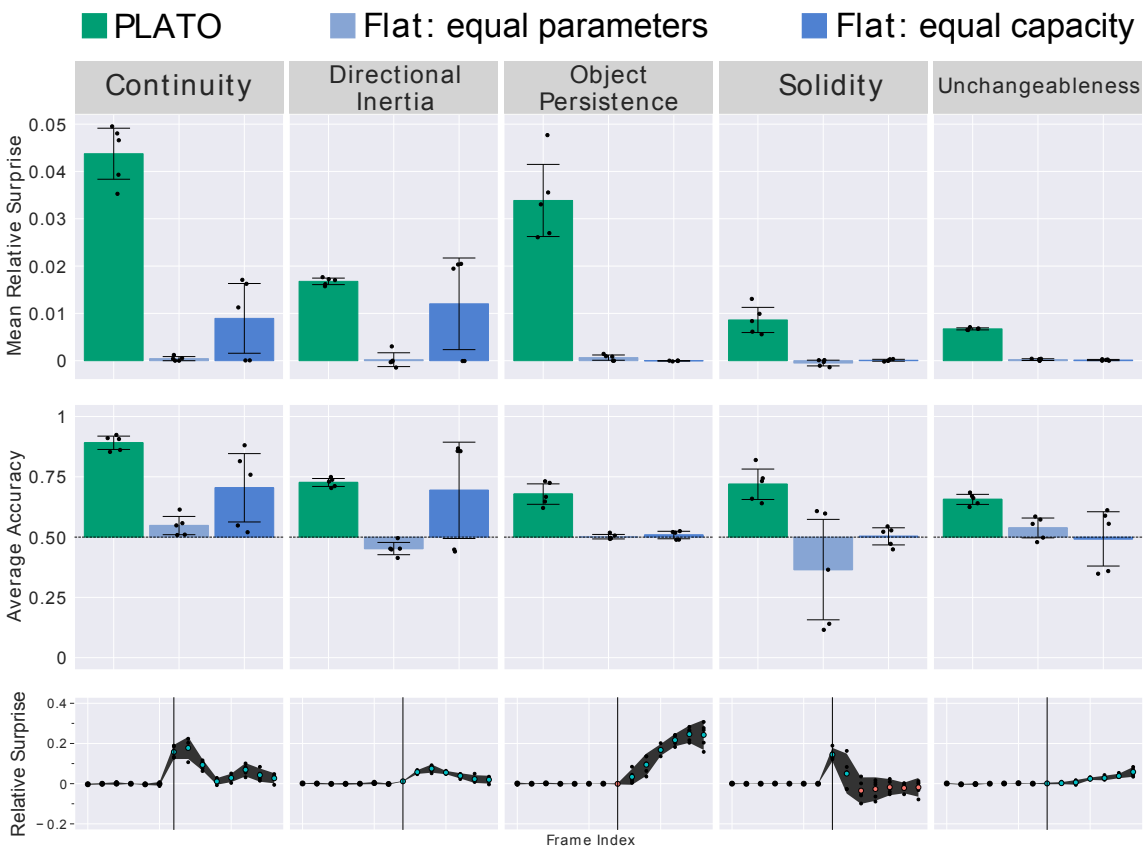


Figure 3.9: **Top, Middle Rows**: Comparison of object-based model, PLATO (green) against flat baseline models (blue). PLATO consistently shows large effect sizes (top row) in response to impossible events and robustly classifies them as impossible (center row). Even when the flat model has more model parameters (dark blue), it only shows above-chance accuracy on two of the five concepts. Each point represents the average performance over 5,000 probe tuples for models trained with a different random seeds. Error bars show the 95% confidence interval over five seeds. **Bottom:** Frame-by-frame analysis of surprise reveals that the relative surprise increases substantially when videos become physically impossible. Relative surprise computed at each frame of probe videos for five random seeds of the PLATO model. The x-axis indicates frame index within a video. Vertical lines indicate when the impossible probes become aphysical (varies by dataset). Blue and red dots code points above or below zero respectively - only applied to frames after the onset of aphysicality. Dots (small) depict different random seeds. Dots (large) depict mean over three seeds with corresponding errorbars showing 95% confidence interval over seeds.

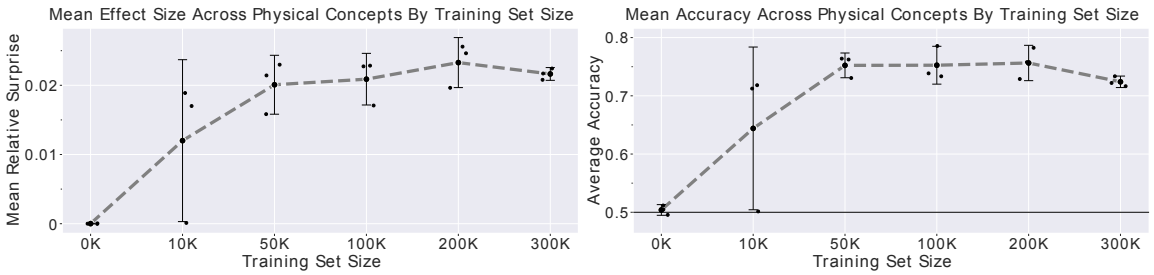


Figure 3.10: PLATO shows robust VOE effects when the dynamics predictor trained with as little as 50,000 examples (28 hours of visual experience). Size = '0K' indicates no training at all and shows that object representations are not sufficient for VOE effects, training the dynamics predictor is critical. Performance was averaged across all five physical concepts (see Supplementary Material for results for broken down by individual concepts). Dots (small) depict different random seeds. Dots (large) depict mean over three seeds with corresponding errorbars showing 95% confidence interval over seeds.

In contrast, VoE effects were essentially absent for the object-agnostic FEP model (see Figure 3.9, Top, light blue bar). The FEC baseline, matched in representational capacity to PLATO (and thus containing many more free parameters) displayed relatively weak VoE effects in two probe categories (continuity and directional inertia; see Figure 3.9, Top, dark blue bar), but no evident VoE effects in the remaining categories.

Taken together, these results indicate a strong facilitative role for object-level representation in the acquisition of intuitive physics concepts, consistent with the conclusions of developmental psychology literature.

3.2.4 Effect of training set size

The training corpus in the Physical Concepts dataset contains a total of three hundred thousand videos. By a conservative calculation (see Methods), this adds up to approximately fifty-two days worth of continuous visual experience, or about five months worth of awake time for an infant. It is a question of obvious interest, both from an AI and a developmental point of view, how much training data is actually required in order to yield VoE effects at test. To assess this, we trained PLATO's

dynamics predictor on datasets of gradually decreasing size, calculating the mean VoE classification over all probe categories. The results (Figure 3.10, Bottom) indicate that robust VoE effects arise in our model after training with as few as 50,000 examples, the equivalent of 28 hours of visual experience (see Supplement for metrics on individual concepts).

3.3 Discussion

'Common sense' reasoning, in numerous varieties, has been an increasingly pressing objective for AI research. In the present work, we reported progress toward building deep learning systems that construct, based on perceptual experience, knowledge in a specific but fundamental common sense domain: intuitive physics. In doing so, we drew inspiration from developmental psychology on two fronts. First, in order to measure progress systematically and quantitatively, we ported the VoE paradigm from developmental psychology to probe five different physical concepts. Second, in order to build a model capable of learning intuitive physics, we endowed our model with object-centric representation and computation directly inspired by accounts of infant intuitive physics. In order to train that model, we created a training dataset of composable physical interactions yielding complex 3D scenes, the Physical Concepts dataset (also making this freely available online as a resource for other researchers). To test the utility of the object-based approach, we compared our model with tightly-controlled, object-agnostic baselines.

Our results center on three key observations. First, our object-based model displayed robust VoE effects across all five concepts we studied, despite having been trained on video data in which the specific probe events did not occur. Second, consistent with expectations founded on the developmental literature, we found that the VoE effects seen in our model diminished or disappeared in well-matched models that

did not employ object-centered representations. Third, we observed that robust VoE effects could be obtained with as little as 28 hours of visual training data.

Our work relates closely to, but also substantially extends, a number of previous studies in AI and computational cognitive science. Two other groups have built intuitive physics evaluation frameworks specifically using the VoE paradigm. Riochet et al (Riochet et al., 2018), in research conducted in parallel with our own original work in this area (Piloto et al., 2018), presented a dataset for probing a variety of physical concepts, including a subset of those covered in our Physical Concepts dataset. Although the videos comprising this dataset include richer textures and lighting effects than our dataset, the physical events themselves are more restricted in their diversity. Smith and colleagues (Smith et al., 2019) employed a related VoE dataset, still simpler in design, and excluding collision events. Neither of these preceding studies reported a system that successfully acquires specific physical concepts through learning: The model presented by Riochet and colleagues showed VoE effects only in a minority of conceptual categories tested, and the model presented by Smith and colleagues embeds a hand-engineered physics engine rather than learning physics from scratch.

While we have spoken in terms of learning physics, our actual model implementation benefited from access to two sources of privileged information: object masks and object indices allowing consistent placement of object embeddings within the model over timesteps. However, it is important to note that in each of these cases recent research has introduced methods that would allow each of these to be extracted from pure video data without access to privileged information of any kind. In the case of object masks, several recent papers have introduced methods for unsupervised object discovery (see, e.g. (Burgess et al., 2019; Greff et al., 2019)). And (Creswell et al., 2020b,a) have recently reported methods for learning to track objects consistently over time. Integrating these methods into our framework, in order to allow true 'learning from scratch,' represents an appealing goal for next-step development. Of

primary interest to the authors is whether or not an integrated system could learn intuitive physics with just 28 hours of visual experience, despite the added tasks of learning object individuation and tracking.

The advantages of object-level representation observed in the present work echo related findings from other areas of AI research, including question-answering (Santoro et al., 2017b) and reinforcement learning (Zambaldi et al., 2018; Bapst et al., 2019; Veerapaneni et al., 2020; Kipf et al., 2019), where object-level representation has been found to support faster learning and improved transfer. What is the explanation for the benefit of object-level representation in the problem setting we have studied? We speculate that that advantage derives from the role that object-level representation can play as a *regularizer*. In a machine learning context, regularization refers to the inclusion of constraints on a learning process, which bias that process to place greater weight on particular inferences from training data (Mitchell et al., 1997). Regularization, in this sense, is employed to prevent a learning system from 'overfitting' to training data, learning representations that are so closely tied to the specifics of that data that the system fails to identify more general regularities that would allow it to perform correctly on new data at test. A bias toward object-level representation could have this kind of regularizing impact on visual learning, in the sense that it biases the learning system toward a compositional representation of visual events which corresponds to the actual compositional structure of the physical events themselves, resulting in knowledge that generalizes well to new events. While this conjecture is, as we have said, speculative, it aligns with one informal observation from preliminary modeling work not described above. Here, we trained object-based and flat models on videos in which the depicted events were much closer in form to those presented at test in the VoE probes. Under these conditions, the difference in VoE effects between object-based and flat models was considerably smaller than what we observed in the simulations that we have reported in the present paper, which fea-

tured much greater separation between training and testing data. It is in this setting, where a learning system must transfer at test to data that lies reasonably far from any particular training example, that regularization can be decisive. Not only does this regularization have the potential to make an intractable task tractable, this regularization may also make models more data-efficient. Where deep learning models have often been criticized as being too data inefficient (Lake et al., 2017), our results demonstrate a relatively small amount of visual experience — on the order of tens of hours — is sufficient to engender robust VoE effects in response to violations of physics. Exploring this regularization-based interpretation of the role of object-based representation thus stands as an appealing hypothesis for further investigation.

Throughout the present work, we have emphasized the role of insights from developmental psychology in guiding the research we have reported. What, if any, are the implications of our work for developmental psychology itself? This topic must be approached with some care, since the model that we have presented is not intended to provide a direct model of physical concept acquisition in children. Nevertheless, we do think there are several insights that may be germane to developmental science. First, our modeling work provides a proof-of-concept demonstration that at least some central concepts in intuitive physics can be acquired through visual learning. Although research in some precocial species suggests that certain basic physical concepts can be present from birth (Chiandetti and Vallortigara, 2011), in humans the data suggests that intuitive physics knowledge develops gradually over time (Spelke, 1994) and can be impacted by visual experience (Wang and Baillargeon, 2008; Smith and Slone, 2017). Of course there is extensive debate and legitimate uncertainty about innateness (Elman et al., 1998). Our modeling work complements these conclusions from experimental research by demonstrating the sufficiency of visual learning to explain the emergence of one core set of physical concepts, assuming — as developmental psychology led us to anticipate — that object-level representations are available.

Furthermore, our model reflects just one point in a family of possible models that implement proposed algorithmic-level (Marr, 1982) principles (e.g. tracked objects) suggested by developmental psychologists. Exploring different ways of implementing these principles (e.g. replacing our object buffer with a recurrent object tracker or removing recurrence from the dynamics predictor (see Supplement)) presents a natural avenue for future work to make more precise contact with developmental accounts at the implementation level (Marr, 1982).

It would be fascinating to extend the present modeling work in order to make even more direct contact with key questions in developmental psychology. For example, the order of concept acquisition throughout development has been of central interest to developmental psychology (Baillargeon, 2004). Although we do not model developmental trajectories in the present work, there is a long history of using connectionist models to shed light on stage-like developmental processes, including in the domain of intuitive physics (McClelland, 1988). Bringing the present work to bear on these questions poses an obvious research opportunity. This would be especially promising in combination with newly emerging datasets that capture first-person footage of infant experience at large scale (Sullivan et al., 2020; Clerkin et al., 2017b). This represents just one opportunity that our model might offer as a tool for computational research into the origins of intuitive physics in human development.

It's worth stating that as interesting as this work may be from the perspective of infant intuitive physics, closely related models (graph networks) are being deployed in AI across a wide range of domains (Battaglia et al., 2018). PLATO's key inductive bias is domain-agnostic: our model represents, computes, and predicts under the constraint that inputs are decomposed into a set of interacting entities. This is essentially the relational inductive bias described in (Battaglia et al., 2018). Given the failure of current AI systems without this inductive bias to learn to perform this obviously helpful (Burgess et al., 2019) discretization, we speculate that the

relational inductive bias may represent a fundamental inductive bias that enables efficient human learning across a range of domains. This would imply that PLATO-like models should be able to learn, where other models fail, in other putatively innate (Spelke and Kinzler, 2007) domains (e.g. agents and their actions). We leave it for future work to catalogue which domains of human reasoning become tractable when a learning system is given the relational inductive bias.

3.4 Materials and Methods

3.4.1 Training Dataset: 'Freeform' Physical Events

Our training dataset consists of 300,000 procedurally generated scenes using custom python code and the Mujoco physics engine. Our motivation was to build a dataset for training that encompassed a wide range of complex physical interactions. Thus, we built composable scene building blocks: rolling, collisions along the ground plane, collisions from throwing or dropping an object, occlusions (via a "curtain" that descends from the top of the screen and retracts), object stacks, covering interactions (an open-bottom, closed-top container falls onto an object), containment events (an object falls into an open-top container), and rolling up/down ramps. Each building block used randomly generated values for object appearances and locations, while ensuring the main intended physical interaction still occurred.

To form a scene, we chose two to four "scene building blocks". Building blocks compose by advertising and/or consuming locations of interest. For example, *rolling* building blocks have two locations of interest: the rolling object's starting point and the point it rolls towards. An *object stack* building block has a single location of interest: the center of the object stack. Composing these two building blocks can yield a scene where a sphere rolls at the object stack. Alternatively, the sphere can

start off at the top of the object stack and roll off of it. With some probability, building blocks do not compose and inhabit the same scene independently.

We restrict the primitive shapes in our dataset to rectangular prisms and spheres. From the rectangular prisms we build a "curtain," a ramp, an arch, and both open-top and closed-top containers. Object sizes were chosen in a hand-crafted range to ensure that objects were visible. Object colors and the checkerboard floor were chosen as random RGB values. Objects had varying but stereotyped masses. Rolled objects and objects in an object stack had a mass of 10. Dropped or thrown objects were made 4 times heavier so that they could more easily displace objects they hit. Containers had a mass of 4 or 5. Arches had a total mass of 60 to keep them upright.

Each scene unfolded over 3000 simulation steps and was rendered into 15 frames at a resolution of 64x64 RGB pixels. Each frame had a corresponding segmentation mask rendered at the same resolution where each object in the scene, including the floor, was given a unique and consistent value. To add extra complexity, we used a drifting camera. The position and orientation of the camera was the same at the start of all scenes. From there the camera randomly drifted its position and location (subject to a bounding box to ensure the camera was still looking at relevant parts of the scene). For each of the 15 frames in a video, we exported the camera position and orientation and gave this information to our model.

Finally, we built an additional 5,000 scenes with this same structure to use as a test set.

3.4.2 Violation Of Expectation (VoE) Paradigm

As described earlier, to assess the model's knowledge of specific physical concepts, we leverage the VoE paradigm from the developmental literature and adapted for artificial models by (Piloto et al., 2018; Riochet et al., 2018) in parallel and subsequently (Smith et al., 2019). We generate 5,000 probe scenarios for each of the following phys-

ical concepts adapted from developmental psychology: object persistence, continuity, 'unchangeableness', directional inertia, and solidity (described below). Whereas the training dataset had a freely moving camera, we used a stationary camera for the probes in order to ensure objects were (where applicable) occluded during the transition from physical to aphysical frames. We compute surprise as the model's prediction errors over the course of a video. Although we optimize the loss in the object code space, for evaluation we look at the error in pixel space. We can take this error in pixel space by decoding object codes to images using the *ComponentVAE*'s decoder, Θ.

$$surprise(video) = \sum_{t=1}^{T-1} \sum_{k=1}^{K} (\Theta(z_k^{t+1}) - \Theta(\hat{z}_k^{t+1}))^2$$

We take this loss in pixel space, instead of in object code space which we used for training, due to the high variance of object codes representing empty "objects." This arises due to the fact that the ComponentVAE emits a fixed number (K = 8) of output components ("objects"). If there are fewer than eight objects in the image, then the ComponentVAE outputs "objects" which correspond to blank pixels. As these empty objects are fairly prevalent in our scene (most videos do not contain the maximum of 8 objects in a scene), the learned representational code uses the unit Gaussian (mean = 0, variance = 1) to represent empty components. To represent components with actual objects in them, the ComponentVAE shifts the mean from 0 and decreases the variance (see Supplementary materials). Because of the high variance on empty objects, it is relatively difficult to predict the exact value of an empty slot across time steps. By taking pixel loss, we avoid this issue as any of the values that encode an empty object all decode to a blank image. Furthermore, because the possible and impossible videos within a probe scenario contain the same objects, we avoid issues that typically arise when using pixel loss. For example, in pixel space, predictions of large objects carry more weight than predictions of small objects, but within a probe scenario the objects in the possible and impossible videos are perfectly matched.

3.4.3 Physical Concept: 'Object Persistence'

Perhaps the most fundamental aspect of intuitive physics is understanding that objects cannot disappear from existence, and is often called 'object persistence' (or 'permanence'). Taking inspiration from a classic behavioral experiment (Baillargeon et al., 1985a), probes for this category involve a rigid plank falling on an object. In the possible probe (Figure 3.2), when the plank falls on the object, the plank occludes it while also remaining propped up by it as expected. By contrast, in the impossible probe the plank falls on top of the object (in a manner that is initially identical) but ends up flat on the floor, as if the object had disappeared. In the counterbalanced probes, the possible probe has the plank falling flat on the floor in an otherwise empty scene, and the impossible probe has the plank falling in the same empty scene but ends up inexplicably propped up by an item that is made to appear under the plank while it occludes part of the floor.

3.4.4 Physical Concept: 'Unchangeableness'

By the principle of 'Unchangeableness' (Baillargeon and Carey, 2012a), objects tend to retain their features (e.g., color, shape) over time. In the possible probes of this dataset (Figure 3.4), a random assortment of static objects are aligned in the foreground. A screen is lowered in front of those objects, and is then raised.

The concept of 'unchangeableness' relates to a number of different aspects of objects, and therefore in the impossible probes we swap the positions of objects when they are behind the curtain to suggest their position, color, or shape has changed.

3.4.5 Physical Concept: 'Continuity'

The concept that an object traces out one continuous path through space and time is referred to in the developmental literature as 'continuity' (Spelke, 1994). For videos

(Figure 3.1) in this category we use a nearly identical setup to a classic experiment (Aguiar and Baillargeon, 1999) where possible probes begin with two static pillars separated by a gap. A ball is rolled horizontally behind both pillars so that it is visible before, between, and after the pillars during its trajectory. In the impossible probes, while the ball is occluded by the first pillar, the ball is made invisible for the period when it would be between the pillars, and then reappears after the second pillar. Alternatively, we make the ball only visible when it is rolling between the pillars, but not before or after. Note: the splicing procedure for probes for this concept deviates slightly from what is shown in Figure 3.8. However, we still maintain the image-level and pairwise matching of frames across possible and impossible videos in a probe tuple.

3.4.6 Physical Concept: 'Solidity'

This dataset is a recreation of an experiment that uses an object and an occluder (Hespos and Baillargeon, 2001c) to test understanding of the solidity of objects, as related to the penetration of an object through a container and the ground below. In probes, perspective is carefully controlled such that the camera can view inside the top of the container but not the bottom. In possible probes (Figure 3.3) a rectangular block is dropped into the container and comes to rest as expected. In the impossible probes, the object "falls through" the container and the floor, and therefore disappears from view (with the penetration itself occluded by the face of the container) even if the object should remain visible due to its height. Alternatively, the object remains visible when its height clearly dictates it should fall further into the container. As this involves judging the relative heights of the object and the occluder from memory, we expect this to be quite a challenging dataset.

3.4.7 Physical Concept: 'Directional Inertia'

This dataset is an adaptation of a paradigm developed by Spelke et al (Spelke et al., 1994a) for investigating infant knowledge of inertia. Where the classic paradigm investigated both magnitude and directional violations of the principle of inertia, the current probe only tests the directional component. Possible probes (Figure 3.5) were formed by rolling a ball at an angle towards a heavy block. Upon contact, the sphere rolled away from the block while reflecting it's velocity about the angle of incidence (as expected). The paired possible probe reversed this trajectory: it began at the end point of the first possible probe, rolled at the same location on the block, and bounced off the block to end up at the initial position of the first probe. Each impossible probe was formed by swapping the trajectories of the impossible probes at the point they make contact with the block. The effect of this swap was that when the ball hits the block, instead of reflecting about the angle of incidence, it heads back towards its initial location, clearly violating the principles of directional inertia for colliding objects.

3.4.8 Model Architecture

To implement our object-centric approach, our model uses two main components: a feedforward perceptual module and a recurrent dynamics predictor.

The perceptual module takes as input a 64x64 RGB image and 64x64 segmentation mask consisting of n=8 channels and produces an object code for each (potentially empty) channel in the segmentation mask through unsupervised representation learning. Although not provided as inputs or targets to the perceptual module, we find that the axes of the code often come to represent interpretable, static properties of the objects such as position on the ground plane, position off the ground plane, color, height or width, rotation, etc (see Supplemental Material for traversals of this learned latent space). These object codes are fed into the dynamics predictor which

is then tasked with predicting the object codes at the next time step. Instead of directly predicting the next image in the video, the dynamics model predicts per-object percepts.

3.4.9 Perception Module

To learn these object codes, we use a ComponentVAE from (Burgess et al., 2019). We used a convolutional neural network for the encoder, Φ, and a spatial broadcast decoder, Θ, from (Watters et al., 2019). Where that work learns segmentation masks using an attention network, we use ground truth segmentation masks from the dataset. For a given image, x, and a segmentation mask, m, with $K = 8$ channels (representing a maximum of eight objects in any scene), the ComponentVAE yields K, 16-dimensional, Gaussian posterior distributions, $q_\phi(z_k|x_k, m_k)$ (where x_k is the masked image). The sample from this code, z_k, is the object code for the kth object. The decoder takes as input z_k to reconstruct masked object image and mask: $p_\theta(\hat{x}_k, \hat{m}_k|z_k)$. Note that like (Burgess et al., 2019), we treat the ground plane as its own object with its own set of parameters in the encoder and decoder. Additionally, we always output eight codes that are used by the dynamics predictor. Where there are fewer than eight objects in the image (some of the mask channels are empty), the "extra" slots are tasked with reconstructing an empty object image and mask.

During an offline phase, before training the dynamics predictor, we optimize the variational objective as defined by (Burgess et al., 2019) to learn the encoder and decoder parameters. We set β to 0.1 and γ to 10 to ensure the model reconstructed the object masks with high fidelity. We found that even at this relatively low value of β, the model learned disentangled representations (Higgins et al., 2016) for each object, corresponding to factors such as position, shape, color, and size (see Supplemental).

We use the pre-trained ComponentVAE in two ways during the prediction task. First, to actually obtain the object codes, we feed in a segmented image into the

ComponentVAE encoder to get the per-object posterior distributions. The object code is then obtained by taking a sample from each posterior distribution. Although it may seem more natural to use the mean instead of a sample, we found better results using the sample on our training data. Thus, each object code is a 16-dimensional vector in a learned representational space. Second, we can use the ComponentVAE's decoder to map the object codes predicted by our model to per-object images (which are easily composed to form a composite image). This allows us to visualize the model's predictions.

3.4.10 Dynamics Predictor

Where the perception module captures static object properties, our dynamics predictor needs to integrate information over time (e.g. to calculate velocities or remember occluded objects). Thus, we employ a recurrent module for the dynamics predictor. We developed the *ComponentLSTM*. The *ComponentLSTM* is an object-wise LSTM (with 2056 hidden units) with shared weights, but object-specific activations. At each timestep, the LSTM for the kth object computes the following:

$$\hat{z}^k_{t+1}, cell^k_t, hidden^k_t = LSTM(z^k_{1\leq t}, cell^k_{t-1}, hidden^k_{t-1})$$

where the $LSTM$ function is shorthand for the standard LSTM(Hochreiter and Schmidhuber, 1997b) updates and $\hat{z}^k_{t+1}$ is the model's prediction for object code k in the next image. The $z^k_{1\leq t}$ here is the *object buffer*: it is the set of all previously seen object codes for the kth object. Although in principle we could feed in just the the most recent object codes, we found much better results on our training data when using the full history. Because the segmentation masks are consistently ordered throughout the course of a video, the perceptual module's outputs are also in a consistent ordering. Thus, when we aggregate object codes into the object buffer

and feed them into the corresponding slot in the LSTM to make a prediction for the corresponding object, we endow our model with *object tracking*.

Thus far, we have not equipped the dynamics predictor with a mechanism for computing the influence of objects on each other. To do so, we leverage an Interaction Network (Battaglia et al., 2016). We compute pairwise interactions *from* the LSTM cell states to both the cell states *and* the object buffer inputs. For the kth memory slot in the LSTM, $cell_{t-1}^{k}$, we compute:

$$int_k = InteractionNetwork(from = cell_{t-1}^{k}, to = [cell_{t-1}^{1:K}; z_{1\leq t}^{1:K}])$$

To aggregate the interactions across all pairs for the kth slot, we compute the sum and the max of the interactions and concatenate these values together. We add the corresponding interactions as an additional input to the kth LSTM, yielding the following update computation:

$$\hat{z}_{t+1}^{k}, cell_t^k, hidden_t^k = LSTM(z_{1\leq t}^{k}, cell_{t-1}^{k}, hidden_{t-1}^{k}, int_k)$$

We call this combined network architecture an $InteractionLSTM$. Additionally, we include a residual connection from the kth object code for the current time step z_t^k to the model's prediction for the kth object (once again exploiting aligned/tracked objects).

Finally, recall that the training data includes a randomly drifting camera. Thus, we feed the $InteractionLSTM$ with viewpoint information as additional inputs. First, the interaction network receives information about the *current* camera position. Additionally, the *ComponentLSTM* receives information about the camera position at the *next* time step. In this way, the dynamics predictor is not only tasked with predicting dynamics, but predicting dynamics from a specific viewpoint.

We train the dynamics predictor as a next-step predictor on the sequence of $T = 15$ segmented images for each video using teacher-forcing. At each timestep, t, we convert an input segmented image to a set of object codes, $z_t^{1:K}$ using the perceptual module's encoder. We form the object buffer, $z_{1\leq t}^{1:K}$, by concatenating these object codes for all previously seen images from the video. To form a fixed-length vector across the sequence, we pad the object histories with zeros where necessary. Before feeding the object buffer into any downstream computations, we project them into a 1680-dimensional space using the Exponential Linear Unit activation function (Clevert et al., 2015). Camera information and the object buffer are fed into the *InteractionLSTM* to yield object code predictions, $\hat{z}_{t+1}^{1:K}$. We optimize the sum of the mean-squared error between our predictions and the observed object codes at the next time step:

$$\mathcal{L}_{object\ code} = \sum_{t=1}^{T-1} \sum_{k=1}^{K} (z_{t+1}^k - \hat{z}_{t+1}^k)^2$$

3.4.11 Baseline Models

Where PLATO used a slotted perceptual module – each slot encoded a distinct object – the flat baseline models encoded the entire scene in a single slot. This translated naturally into the dynamics predictor which made predictions based off of a single slot. The updates in the flat models are identical to the PLATO update equations above with the exception that $K = 1$ instead of $K = 8$. Because there is only one slot, this means the flat module does not get any information in the segmentation mask (it has only a single channel and is uniformly filled to 1). However, see (Riochet et al., 2018) for evidence that non-object based models trained with segmentation masks still fail intuitive physics examinations.

3.4.12 Experiments

Training To train the perception module, we used the RMSProp optimizer, with a learning rate of 1e-4 for 1,000,000 steps with a batch size of 64 images. To train the dynamics predictor, we trained our models for a total of 1,300,000 training steps with a batch size of 128 videos. We used a learning rate that transitioned from 1e-4 to 4e-5 after 300,000 training steps. The above training procedure was used for all models and all experiments.

PLATO Vs. Flat Models For each model, we pretrained the perception module on the full dataset of 4,500,000 images (300,000 videos each with 15 frames of content). We then trained the dynamics predictor, freezing the weights of the perception module, using five different random seeds for weight initialization. For each seed of the dynamics predictor, we ran 5,000 probe quadruplets for each of the five concepts.

Training Set Size As above, we pretrained a perceptual module on the full dataset size of 300,00 videos (4,500,000 images). We then varied the dataset size used to train the dynamics predictor using three random seeds while freezing the weights of the perception module. Training the dynamics predictor proceeded as above regardless of dataset size. We also included a '0K' training set size which corresponded to a model with an untrained dynamics predictor. This model still received object percepts from the pretrained perceptual module. The lack of VOE effects for the '0K' model shows that the VOE effects depend on more than just tracked object representations. Predicting dynamics is necessary for VOE effects on the five physical concepts probed.

3.4.13 Visual Experience Calculation

To compute the visual experience contained in our dataset, we calculated the following: 300,000 videos each lasting two seconds yielded 600,000 seconds or roughly 6.95 days of continuous visual experience. For only 50,000 videos, this amounts to roughly

28 hours of continuous experience. Conservatively assuming eight hours of wakeful experience in a day yields an equivalent of roughly **20.9 days** for 300,000 videos or **3.5 days** for 50,000 videos.

4

Object Representation and Tracking

Our initial plan for PLATO was to learn intuitive physics in a completely unsupervised fashion. Instead, PLATO relies on tracked object masks. To learn this, we would have to build an object decomposition model to produce the masks and an object tracking module to keep track of objects as they progressed through the scene. Unfortunately, the most promising method available at the time simply did not work well when applied to our datasets. Objects were frequently over-segmented based on color. Before resorting to using ground-truth, tracked object masks for PLATO, we attempted to improve unsupervised object decomposition for use in a model that learned intuitive physics. The current chapter outlines the work done towards that goal. While our methods showed promise in certain domains, they nonetheless failed to achieve satisfactory results on our intuitive physics dataset. Despite being novel research approaches and problems, the work in this chapter was done with the specific goal to use in our intuitive physics model. As such, the majority of the results in this chapter are brief and qualitative in nature: we evaluated our techniques based on whether or not they would work on our intuitive physics dataset.

Unlike the previous chapters, this work was never written up outside of the current thesis. As such, we acknowledge here the contributions of Antonia Creswell as a co-lead on the version of the AlignNet described here. Antonia went on to develop the AlignNet work even further (see Creswell et al., 2020b,a).

4.1 Slow Multi-object Network

4.1.1 Introduction

Our deep neural network model from Chapter 3, PLATO, was decidedly object-centric in its representation and computations. We could assure relational computations simply due to how we constructed the network's connectivity: we used an Interaction Network (Battaglia et al., 2016) in our dynamics model. However, giving the model object representations (which would be computed over in our dynamics module) required two steps. First, we had to allow for discrete entities. This was easily achieved by sculpting the model's architecture to route the output of perception into a set of discrete slots. The more difficult step was to actually get representations of objects into the slots. This is the object discovery/segmentation/representation problem described in Chapter **??**, which remains an open issue. One way to achieve this with learning was to apply recent advances in unsupervised object discovery (e.g. Burgess et al. (2019)) to our dataset as a perceptual front-end. This front-end would take an input image and produce a fixed-size set of vectors, where each vector represents an individual object in the image (or no object if there are fewer objects in the scene than the set size). When applying MONet (Burgess et al., 2019) to our dataset, we found it frequently over-segmented objects based on color. It seemed that the model found ways to efficiently decompose a scene into "motifs" that just were not aligned to what we would describe as an object.

Even assuming MONet could successfully discover objects in our dataset, there is an additional issue that arises when using sets of objects for next-step prediction. If we make predictions about a future timestep that consists of a set of objects, then we introduce an alignment problem: which of our model's predicted set of objects correspond to which objects in our set of targets? One approach is to assign an alignment that minimizes the error for all predictions. In practice, we observed in prior experiments that this naive solution did not work well for learning. Our models got stuck in sub-optimal local minima where predictions tended to correspond to the average of all object features. An alternative solution is to implement object tracking. If we impose an ordering over the set of objects, then our model can always assume that it should place its prediction for an object in the slot index where it first observed that object. This eliminates the alignment problem at the time of prediction and instead handles it upfront. Of course, the question becomes how can we implement object-tracking through learning?

In the present work, we extended MONet to the video domain using a recurrent architecture, which we called SlowMONet. We hypothesized that incorporating motion information alongside a "slowness" regularization term would improve MONet's object discovery performance and allow it to stably decompose objects into an ordered set across time. This approach has the potential to solve the object discovery and tracking problems simultaneously. We augmented MONet's objective function with a slowness term. This slowness term penalized large changes in the representation in a specific slot across time steps. This was motivated by the observation that most features of an object do not change throughout the coarse of a video and, therefore, the representation of an object at the current timestep should look quite similar to the representation at the next timestep. Furthermore, when object properties do change over time (e.g. position for moving objects, color due to lighting changes, size due to moving closer or further from a camera), they tend to change gradually. This

has the effect of keeping objects in the same slot throughout the course of a video. If they changed slots randomly from frame to frame, the change in representation would be large and the slowness penalty would be triggered. By adding recurrence across frames and the slowness prior, we hoped to augment MONet's unsupervised object decomposition to be motion-sensitive and stable (i.e. object tracking).

4.1.2 Methods

SlowMONet is an extension of an early incarnation of the MONet model. We briefly review the MONet architecture and subsequently we describe our alterations.

MONet

MONet is designed for unsupervised scene decomposition of single images. It does this by splitting an image into K components, individually reconstructing each component, and combining all K reconstructed components to reproduce the original image. To do this, MONet uses an attention network and a *ComponentVAE*.

The attention network is a recurrent U-net (Ronneberger et al., 2015) neural network which runs over $K-1$ steps. It takes two inputs: an image x and a "scope", s_k The scope is the same height and width as the image to be decomposed. It can only take values in the range $[0, 1]$. Intuitively, it can be thought of as what still needs to be explained in the original image. In the first pass of the attention network, the scope (s_0) is uniformly set to 1, indicating that the entire image needs to be explained (segmented). The output of the attention network at the kth application is an attention mask, m_k which again takes values in the range $[0, 1]$ and has the same height and width as the input image. This attention mask is directly interpretable as the segmentation mask for the kth object (alternatively called a component). That is, every pixel in the attention mask that is set to 1 in the kth attention mask is interpreted as belonging to the kth object. The scope, to be used in the next

application of the attention network, is updated to "remove" the pixels explained by the current attention mask: $s_{k+1} = (1 - m_k) \odot s_k$. The attention network is applied in this way a total of $K - 1$ times each time removing from the scope what has not been explained by previous passes of the attention network. The final attention mask, m_K, is not outputted by the attention network. Instead, it is taken to be the final scope value, s_{K-1}. In essence this sets the final attention mask to anything that has not yet been explained. This ensures that the entire image is explained and the network cannot find a degenerate scene decomposition which is easy to explain because it simply chooses never to explain anything at all.

The attention masks are used to indicate how the ComponentVAE should reconstruct the entire image. Each attention mask, along with the original input image, is passed independently through a variational auto-encoder (VAE). This VAE uses a convolutional neural network to yield a latent posterior distribution, z_k, for the kth component. As typical for a VAE, a sample is then drawn from z_k and passed through a decoder network. MONet uses a spatial broadcast decoder (Watters et al., 2019) to produce reconstructions of the image and the attention mask.

The key mechanism in MONet is to mask the standard reconstruction loss of the VAE objective by the attention mask. The kth pass through the VAE is only penalized for its reconstruction in the pixels that were selected via the kth attention mask. In this way, the attention network directly controls the reconstruction loss in each pass through the VAE. Besides the masked reconstruction loss, MONet uses two additional loss terms.

The second term is standard to the VAE framework: it is the Kullback-Leibler divergence (D_{KL}) of the latent posterior, $q(z_k)$, with the latent prior, $p(z_k)$. The KL-divergence is weighted by a hyperparameter, β, in accordance with the β-VAE framework (Higgins et al., 2016).

Finally, the third term is the reconstruction of the attention mask. This is kept separate from the image reconstruction loss because it should not be masked. Additionally, it is given its own hyperparameter, γ. The masked image reconstruction loss, factorized, per-component KL-divergence, and mask reconstruction losses can be succinctly written over all components in the following form, where ϕ and θ correspond to the parameters (shared across all components) of the encoder and decoder networks respectively and psi corresponds to the parameters of the attention network.

$$\mathcal{L}(\phi;\theta;\psi;x) = -\log \sum_{k=1}^{K} m_k p_\theta(x|z_k) + \beta D_{KL}(\prod_{k=1}^{K} q_\phi(z_k|x,m_k)||p(z)) \tag{4.1}$$

$$+ \gamma D_{KL} q_\psi(c|x) p_\theta(c|z_k) \tag{4.2}$$

Before describing our modifications, we want to give insight into the motivation of MONet. The KL-divergence between the posterior and prior distributions, in addition with the reconstruction loss, can be interpreted in VAEs as instantiating an information bottleneck (Tishby et al., 2000): a constrained optimization problem where the objective is to maximize the mutual information between the reconstruction targets and the latent representation while minimizing the mutual information between the latent representation and the input. The information bottleneck principle, when used in an autoencoder, can be thought of as a dimensionality reduction technique (e.g. principal component analysis) with an information-theoretic regularizer which encourages compression. In this way, the network is encouraged to learn compressive representations of *just* the pixels represented in the attention mask, while discarding any information about the rest of the image (i.e. the pixels that are not in the attention mask). This compressive pressure is critical in pushing MONet towards discovering objects. To understand this, let us consider various attention mask proposals over an imagined dataset of images (Burgess et al. (2019) make this argument

in more detail). It will consist of a 3D ball placed in various scenes consisting of many shapes and colors. The ball's surface is a single color, although may occupy a range of pixel values due to lighting. In one case, let's consider an attention network that perfectly attends to the pixels in the ball. In another case, let's consider an attention network that outputs an attention mask of the same size, but in a random location in the image. In the first case, the entropy of the attended pixels is fairly low. Given one pixel in the ball, we are likely to know that other pixels in the attended mask have a similar color. In the second case, knowing the value of one pixel in the attended region is much less (if at all) informative about the other pixels in the attended mask. The attention mask could contain multiple objects. Burgess et al. (2019) make the argument that attention masks that capture individual objects have less entropy, they allow more compression. So if an attention network must learn to carve up an image for optimal compression, then an effective strategy is to learn an attention mask for each object. It is through this combination of architecture and optimization objective that the authors seek to learn unsupervised object discovery.

SlowMONet

Where MONet capitalizes on the fact that pixels within the same object share information within an image, SlowMONet exploits the shared information of objects across frames of a video. If I observe a red ball in the previous frame, then I can substantially reduce my uncertainty about its properties (e.g. color, size, location, etc) in the current frame (assuming, of course, reasonable temporal resolution). Let's consider this implication in terms of the outputs of MONet. For a video sequence, X, consider the application of MONet to two adjacent image frames, x_{t-1} and x_t. This would yield, amongst other things, two sets of attention masks, $m^{t-1}_{1:K}$ and $m^{t}_{1:K}$, and two sets of latent representations, $z^{t-1}_{1:K}$ and $z^{t}_{1:K}$. If object properties are highly correlated across time, then for a learned scene decomposition that successfully captures

objects, we would expect that the latent representation for some object, b, should be similar across two frames: $z_{i_b}^{t-1} \approx z_{i_b}^{t}$. We use i_b to indicate the component index for object b for the current timestep. MONet does not impart any consistency between the ordering of decompositions across repeated applications: an object may be selected as the first component at timestep $t-1$ and the last component at timestep t.

The key modification in SlowMONet is to explicitly enforce $z_k^{t-1} \approx z_k^t$. This is done by changing the prior distribution over latents. In MONet, the prior distribution, $p(z_k)$, is a standard Normal with diagonal covariance. In SlowMONet the prior distribution[1] for the kth component is the inferred latents of the kth component at the previous timestep: $p_t(z_k) = q_\phi^{t-1}(z_k|x, m_k)$. This modification explicitly encodes our assumption that object properties change slowly over time. It is a slowness regularizer. It has direct interpretation as a probabilistic (Turner and Sahani, 2007) slow-feature analysis model (Wiskott and Sejnowski, 2002). By ensuring that representations in the k^{th} component are similar throughout a video, we encourage the model to assign the same component index to an object throughout the video. If a model can succeed in doing this, it has learned object tracking. Without any changes to the MONet architecture, a model could never learn to satisfy our slowness regularizer. That is because MONet operates over single images. How could the attention mask match the ordering of objects from a previous frame without information of the ordering in the previous frame. To fix this, we augment the inputs to the attention network to include the attention masks from the previous timesteps. This yields an additional benefit to SlowMONet: it can use motion to segment and track otherwise ambiguous objects.

[1]For the first timestep where there is no previously inferred latents, we revert to the prior used in MONet.

$$\mathcal{L}^t(\phi;\theta;\psi;x^t) = -\log \sum_{k=1}^{K} m_k^t p_\theta(x^t|z_k^t) \tag{4.3}$$

$$+ \beta D_{KL}(\prod_{k=1}^{K} q_\phi^t(z_k|x^t,m_k)||q_\phi^{t-1}(z_k^{t-1}|x^{t-1},m_k^{t-1})) \tag{4.4}$$

$$+ \gamma D_{KL} q_\psi^t(c|x^t) p_\theta^t(c|z_k^t) \tag{4.5}$$

Datasets

We tested SlowMONet using two datasets which were both rendered in 64x64 pixel image sequences. The first dataset was procedurally generated and consisted of up to 4 squares drifting over a blank background. It consisted of 10 frames. Each square was independently assigned a random color, meaning that within the same video two or more objects could in principle have the same color. For our second dataset, we used the training data from our experiments in Chapter 3. Each video sequence had 15 frames, where there were a maximum of 9 unique objects.

Architecture and Hyperparameters

MONet requires the specification of an attention network and the ComponentVAE's encoder and decoder modules. We adopted a standard U-net, convolutional neural network, and broadcast decoder respectively. For those modules, we used the same architectural choices from (Burgess et al., 2019) with the exception that we used K = 9 slots. For optimization, the RMSProp optimizer with a learning rate of 1e-4 and 500000 training steps with a batch size of 384. We set $\beta = 0.1$ and increased γ throughout training in a step-wise fashion. It started at 0.1, increased to to 0.7 after 20000 steps, and ended at 1 after 40000 training steps.

4.1.3 Results

We first sought to test whether or not MONet was actually using motion information and the slowness penalty to successfully implement object tracking. We developed a simplistic dataset consisting of colored squares drifting within a bounded region of space. We found that SlowMONet was capable of identifying the individual squares as objects (MONet has been observed to be color sensitive, after all). Furthermore, we found that the placement of the objects into MONet's slots was stable throughout videos. To actually test whether or not MONet was using motion for stable object discovery, we evaluated it on videos where all the objects were identical in shape, size, and color. In this case, motion is the *only* cue SlowMONet could have for consistently placing objects in the same slot throughout a video. Figure 4.1 shows that SlowMONet did in fact track the identical objects according to their trajectory of motion. We applied SlowMONet to our intuitive physics dataset and found mixed results. Figure 4.2 shows a cherry-picked example. First, motion was not a strong enough cue to prevent the model from oversegmenting the checkerboard floor pattern into two separate slots. It also over-segments the bottom of the pink container as a separate object - which starts off in one slot (colored white in the far right) and then switches to another slot (colored purple). On the other hand, many objects are in fact correctly decomposed and tracked. Of particular interest is the object falling in the top-left of the images. It is mostly out-of-frame and falling to the ground, but SlowMONet identified it as an object and tracked it throughout its descent. Unfortunately, inspection of SlowMONet on other examples were less successful, which lead us to abandon it for the PLATO project.

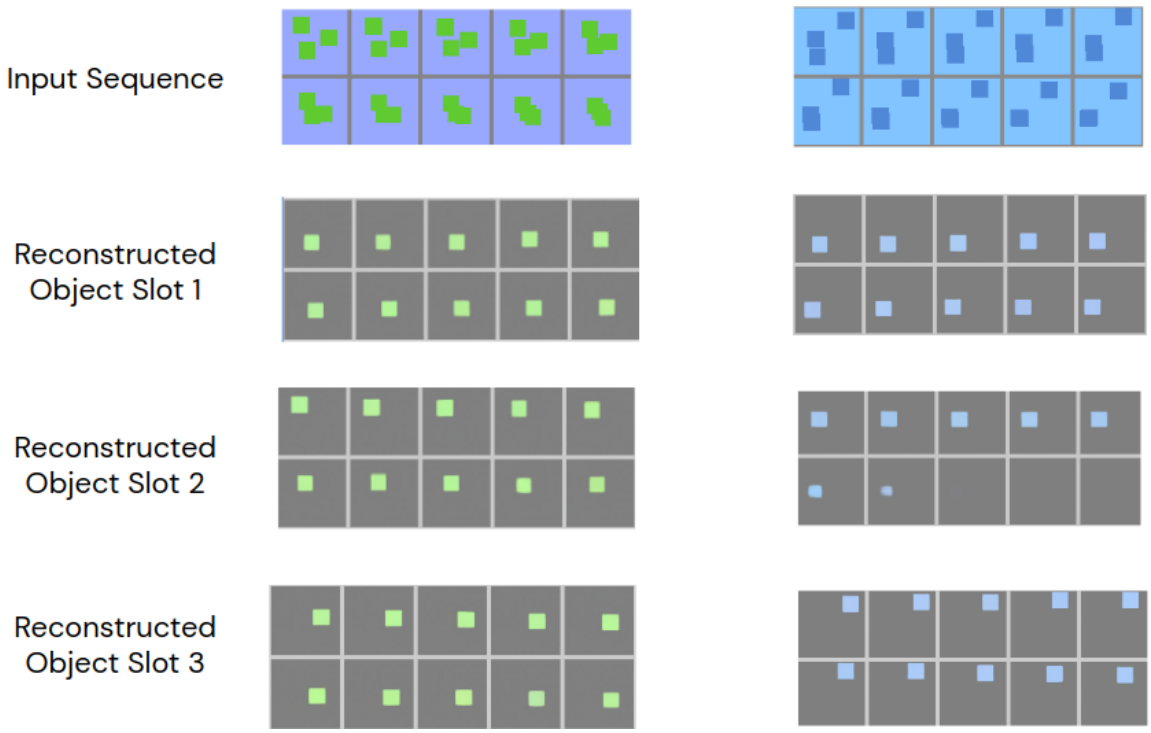


Figure 4.1: SlowMONet successfully uses motion information to track identical objects in a simple multiple object tracking dataset. **Left Column:** The top row shows an image sequence of ten frames, which starts at the top-left going rightward and wraps around after five frames. The sequence consists of three identical green squares drifting within a bounded square region. Below the input sequence we show the reconstructed object for each object slot. We can see that despite having identical visual features, SlowMONet keeps each object in the same object slot throughout the sequence. **Right Column:** The top row shows another input example consisting of identical blue squares. The second object slot reconstruction shows that despite stably tracking all visible objects - the model's representation of perfectly occluded objects does not persist longer than one frame.

4.2 AlignNet

4.2.1 Introduction

In SlowMONet, we sought to perform object decomposition and tracking in a single system. Given only a partial success with this approach, we sought to attack just the object tracking problem assuming an object decomposition was given. Concretely, object tracking is the process by which we observe a set of objects, U_t, and form a correspondence from U_t to objects in the future, U_{t+1}. That is, for each element of



Figure 4.2: SlowMONet applied to Physical Concepts dataset. The sequence of images is shown starting at the top-left and wraps around to the next row. **Left:** The input data depicts two containers (one green, one pink) and a yellow ball. Additionally, a box, that is only ever partially in-view, falls from above the viewport, hits the corner of the green container, and bounces out of frame again. **Center:** SlowMONet's reconstruction of the image sequence. The reconstruction is fairly accurate, but a bit blurry. Some colors (i.e. specular highlight on the ball) are muted. **Right:** The segmentation masks learned by SlowMONet. Each unique color corresponds to an inferred object. The green container (assigned the light blue color slot) is correctly segmented in the same slot throughout the entire sequence. An impressive result from this example is that the falling box (assigned the red color slot) is correctly segmented into a stable slot through the sequence. Errors in segmentation are seen where the model over-segments the checkerboard floor, assigning one object slot to each of the. Errors in the stability of segmentation are seen in the segmentation of the ball which is initially assigned the purple slot but after being completely occluded in the fourth frame emerges again in a different slot (the green slot assigned to part of the checkerboard).

U_t we want to assign an *aligning index* into U_{t+1}. Such an object tracking system is necessary when using a model like *PLATO*, which makes per-object predictions. There needs to be an alignment between the object predictions and the ground truth predictions. There are two sources of difficulty in this version of the object tracking problem.

The Transformation Problem

The transformation problem is that objects may undergo transformations which preserve identity, but perturb their properties. If I shine a light on a previously unil-

luminated ball, it looks different but it is still the same ball. A ball rolling across a screen changes positions but it is still the same ball. If object properties do not change at all across the two sets, alignment does not require learning at all. It can be solved with standard algorithms (i.e. Hungarian matching algorithm) by constructing an adjacency matrix from the pairwise distances of the objects in the two sets. If object properties do change across the two sets to be aligned, solving the transformation problem amounts to learning which object properties can change while retaining identity[2].

The Effective Cardinality Change Problem

The effective cardinality change (ECC) problem is another difficulty for object tracking. It arises when there is an effective change in the cardinality between U_t and U_{t+1}. Object discovery algorithms like MONet Burgess et al. (2019) or IODINE Greff et al. (2019) produce a set of objects with a fixed size, K. When there are only $N < K$ objects in a scene, the remaining $K - N$ set elements contain representations that do not render any pixels at all. They are effectively empty or null, but have the same representational form (i.e. a vector) as non-empty elements. Thus, despite the fixed set size of these models, the *effective set size* can change based on the input image. This can happen when objects go out of view ($|U_t| > |U_{t+1}|$). Or when previously unseen objects come into view ($|U_t| < |U_{t+1}|$). Without a change of cardinality, each object in U_t whose correspondence we want to learn can be assigned a unique index to its object in U_{t+1}. Essentially the assignment matrix is a permutation matrix. With a change in cardinality, valid alignments are no longer bijective. This means we can no longer solve the alignment by trying to match a specific permutation matrix. To see this let's consider the case of *added* objects with *effective* set sizes: $|U_t| = |U_{t+1}| - 2$. In this setting, there are two "empty" objects in U_t which need to get paired with the

[2]In this sense, it seems that learning object tracking is deeply intertwined with learning object dynamics

newly added objects from U_{t+1}. In this case, there are now multiple valid alignments. Either empty object can be correctly aligned to either added object, as long as both added objects are assigned to an empty object and no added object is assigned to more than one empty object. Similar issues arise when objects are dropped.

Our approach was to gradually increase the complexity of the alignment problem. First, could we learn to align the same exact objects that are simply shuffled? This avoids the transformation problem *and* the effective cardinality change problem. As mentioned above, this problem setting does not require learning, but, as our results showed, can be solved with learning. This demonstrated the viability of our proposed *AlignNet* architecture and self-supervised training method. We then evaluated whether or not our model could learn to align objects across time using labelled (supervised) data. This indicated whether or not the model could cope with changes to object properties that might occur over time. This case does invoke the transformation problem. We then addressed the issue of adding or dropping objects in isolation. In this setting, we addressed subparts of the effective cardinality change problem (without the transformation problem). Finally, we upped the stakes by learning to align objects across time with some objects being added or dropped, again using labelled data in this case. This addressed both the transformation and ECC problems. Overall, we found our model was capable of learning to align objects in this regime. We describe our model and results in the next section. Although the work in this section only addressed supervised and self-supervised alignment (and therefore would not be suitable for unsupervised object tracking in PLATO), it set the stage for our colleagues (with whom the current work was done) to work on the fully unsupervised case (see Creswell et al., 2020a).

4.2.2 Methods

General Problem Specification

As described above, we address multiple versions of the alignment problem. We adopt a common notation for all versions. We always consider two sets of objects: U and S. Each object is represented as a vector. Some of these vectors have a semantic interpretation as corresponding to no object. We wish to find *aligning indices*, A, for each element of U, where A_i specifies which index in S is matched to the i^{th} object in U. We refer to these aligning indices as A when they are targets (provided by supervision or self-supervision) and $\bar{A}$ when they are the model's output.

Datasets

We use two datasets in different configurations to address the various versions of the alignment problem. The first dataset is a synthetic symbolic dataset of objects with discrete properties represented as vectors. We refer to it as the *Symbolic Objects* dataset. It consists of configurations of 8 objects, each with the following attributes:

- x-coordinate: the object's x position in a 2D space. It takes 7 distinct values ranging from $[-0.6, 0.6]$, with a step size of 0.2
- y-coordinate: the object's y position in a 2D space. It takes 7 distinct values ranging from $[-0.6, 0.6]$, with a step size of 0.2.
- "R": the red channel in an RGB color space. It takes 5 distinct values ranging from $[0.2, 1]$ with a step size of 0.2.
- "G": the green channel in an RGB color space. It takes 5 distinct values ranging from $[0.2, 1]$ with a step size of 0.2.
- "B": the blue channel in an RGB color space. It takes 5 distinct values ranging from $[0.2, 1]$ with a step size of 0.2.

- shape: the shape of an object. It takes 6 distinct values ranging from $[0, 1]$ with a step size of 0.2.

Despite the fact that this dataset ultimate boils down to individual examples that are just a set of vectors, we found naming the dimensions (e.g. shape) to help build intuition.

Our second dataset was based on the image training data from our experiments in Chapter 3. Recall that the AlignNet operates over sets of vectors (presumably representing objects). In order to convert the images in our dataset into sets of object representations, we leveraged the fact that our training data had ground truth segmentation masks for each object. We trained a variational autoencoder (VAE) on each object in an image as a learned pre-processing step. This follows the same pattern used in PLATO, but we briefly summarize it here as a refresher. For a given image X and a set of segmentation masks $m_{1:K}$, we generate a set of masked images, x_k where only a single object is visible. We run each masked image through a convolutional encoder, $\phi(x_k)$ to produce a latent distribution z_k. We take a sample from this distribution and decode it through a broadcast decoder θ to produce a reconstruction of x_k. We optimize the weights of the encoder and decoder to make accurate reconstructions as is standard for the VAE framework. In this way, we have our *Physical Concepts Objects* dataset which comprises a sequence of $T = 15$ frames. Each frame consists of a set of vectors, $z^t_{1:K}$ representing objects in a learned embedding space. Additionally, we can visualize any object representation, z_k, by running it through the VAE decoder. We chose this dataset partially because it is practically relevant to our research goals for PLATO. But also because it was a more natural test of the AlignNet given that objects were represented in a *learned* space instead of the (intentionally) contrived attributes of the Symbolic Objects dataset.

Finally, we contrast the various affordances of each dataset. Firstly, the Symbolic Objects dataset is atemporal. On the otherhand, the "Learned Objects" dataset can

be digested either as independent frames or by consuming pairs of frames $(t, t+m)$. Leveraging pairs of frames allows us a realistic instantiation of the transformation problem. Another difference is that the effective set size of Symbolic Objects dataset matches the fixed-size of the set. In other words, there are no "empty" objects in that dataset. The Learned Objects dataset on the other hand has many instances of empty object slots. These empty object slots are represented using a code determined by the VAE. As such, we do not explicitly know which object slots are empty in the general case. Furthermore, there is no guarantee that two empty slots have the same vector representation. Another consequence of learning object representations is that we do not know, in the general case, what properties each axis of the representational code correspond to. In the Symbolic Objects dataset, by design we labelled the axes to have semantic values (a property which we later exploit).

4.2.3 AlignNet V1: Learning Similarity

In this section, we describe the first version of our model and framework for learning to align objects. It comprises a fairly simple architecture and a self-supervised training procedure. It does not address the effective cardinality change problem, only the transformation problem. We use a variety of settings to establish when learning is and is not necessary.

Architecture and Optimization

Our actual model was built off of key-query attention networks that have gained prominence (Vaswani et al., 2017). We use a multi-layer perception (MLP), with shared weights, to produce a "query" for each object in U. We use a second MLP with its own weights to produce a "key" for each object in S. In practice, we actually used a third MLP which acted on both the U and S separately before being passed into the respective query and key MLPs. By computing either cosine similarity (or

$l2$-norm) from each query to each key and normalizing across all keys, we can build an attention matrix, $\hat{A}$. This can be interpreted as a probability distribution over how well each object in U matches each object in S according to the similarity in the learned key-query space. We can then take the aligning index for each query as the element with the highest value in the attention matrix for the corresponding object: $\bar{A}_i = argmax(\hat{A}_{i,1:|S|})$. Note that $\hat{A}$ is not guaranteed to be a proper permutation matrix and therefore multiple objects in U can be aligned to the same object in S. To train our model, we modify our network parameters to minimize the summed row-wise cross entropy between the true aligning matrix and the model's predicted alignment:

$$L = -\sum_i \sum_j A_{ij} log(\hat{A_{ij}})$$

All experiments were carried out using the ADAM (Kingma and Ba, 2015) optimizer with a learning rate of $1e-4$ over 10000 training steps. We clipped gradients with absolute values larger than 10 and found this stablized training significantly. The shared "preprocessing" MLP had two layers with 64 and 128 hidden units respectively and used the ReLU activation function. The query and key MLPs also had two layers with 128 hidden units in the first layer and 20 units in the second layer and again used the ReLU activation function.

The Transformation Problem

First, we consider learning to unshuffle the same set of objects. In this case, $S =$ Shuffle(U). In this case, the objects to be aligned are identical. There is no transformation of object properties. There is no benefit for a learning-based approach. A simple algorithm can solve this problem: take the aligning index to be the index which maximizes a similarity metric from objects in U to objects in S. In the coming results, we use this simple algorithm as a baseline using either the (negative of the)

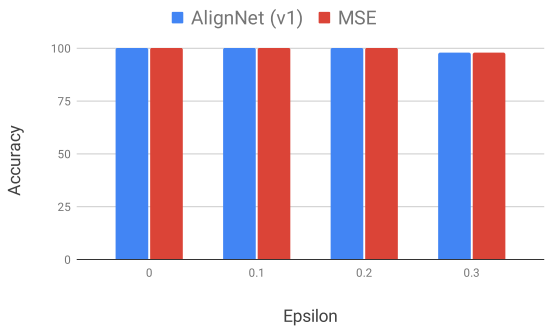


Figure 4.3: Classification accuracy for aligning with various amounts of noise (controlled by ϵ) in the Symbolic Objects dataset. We compare the baseline, no-learning approach "(MSE)" against the AlignNet. Even at the highest noise level, both models achieve high accuracy.

mean-squared error or cosine-similarity as a similarity metric. In either case, the baseline can be interpreted as whether or not learning is required to solve the alignment problem. We show results for our model and the baselines in this simplest of cases along with a family of alignment problems: $S = \text{Shuffle}(\text{UniformNoise}(U, \epsilon))$, where ϵ controls the amount of noise added to objects. Concretely, given that each object is represented as an n-dimensional vector. We perturb each element of this vector with independently sampled random noise. When $\epsilon = 0$ this reverts to our initial formulation where object properties do not change and learning is not necessary. By setting $\epsilon > 0$, we confront the transformation problem. This is the motivation for adding the noise term: it allows us to simulate in a parametric fashion the impact of the transformation problem on alignment performance.

Figure 4.3 shows classification accuracy as we vary the noise level using the Symbolic Objects dataset. As expected, when $\epsilon = 0$ both the AlignNet and the baseline models achieve 100% alignment accuracy. This establishes that at the very least, learning does not hurt alignment in the simplest case. As we increase the value of ϵ we can see the accuracy of all models being to dip. This stands to reason: at some point the magnitude of the noise stamps out any indication of the shuffled object's identity. We include this graph for posterity, but it is rather unsurprising.

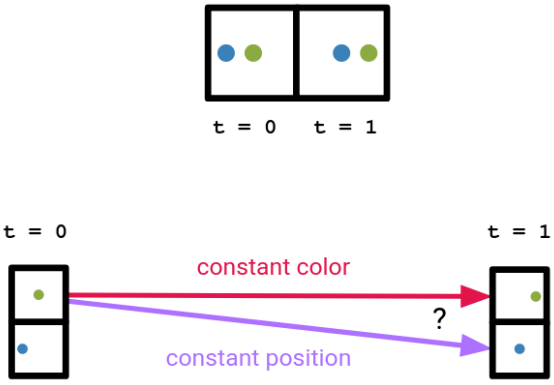


Figure 4.4: **Top:** Illustration of a two-frame sequence with a green ball and a blue ball in each frame. **Bottom**: Without assuming any structure in how objects are allowed to transform while retaining identity, the correct alignment is ambiguous. The red arrow suggests aligning to keep color constant. The purple arrow suggests aligning to keep position constant.

The (Simulated) Structured Transformation Problem

In the previous section, we added uniform random noise to every property of an object. If the world operated this way, it would amount to trying to track objects which can change position, color, size, shape, material, etc. In such a setting, object tracking is at the mercy of the magnitude of the perturbation and no amount of learning will overcome that. Assuming a world where *all* object properties can change, figure 4.4 shows the inherent ambiguity in trying to track objects in a pair of frames. It is impossible to tell whether the green ball from the first frame retained its color but changed its position or retained its position and changed its color. However, the world does have structure to it which can be exploited. Object properties change in principled ways. They tend to change position but not shape or material. To simulate what we will refer to as the *structured transformation problem*, we add uniform random noise to only a subset of object properties. We call this *structured noise*. This simulates the idea that only *some* object properties change. It is in this experimental setting where the AlignNet has the potential to outperform the baselines. It can learn to put higher priority on the constant object properties instead of weighting all of

them equally. Figure 4.5 shows the classification accuracy as we vary ϵ for structured noise on the Symbolic Objects dataset. We add noise to the first two dimensions to simulate the idea that object positions can change, but other features (e.g. like shape) do not. The AlignNet achieves constant, near-perfect accuracy even as we increase ϵ to values which lead to significant performance declines for the baseline model.

Figure 4.6 shows a similar result when we apply structured noise to the Learned Objects dataset. One aspect of this figure that requires clarification is that accuracy is not at 100% for the case where $\epsilon = 0$. This is due to the fact that this dataset has empty objects. If an example has multiple empty objects in it (as frequently occurs in this dataset), then there are multiple valid alignments from the shuffled empty slots to the unshuffled empty slots. Although a theoretical concern, in the present work we ignore this and simply use the value at $\epsilon = 0$ as a reference for interpreting our results. Just as in the previous result, we can see that the AlignNet's accuracy remains nearly constant as we increase ϵ while the baseline model accuracy begins to drop.

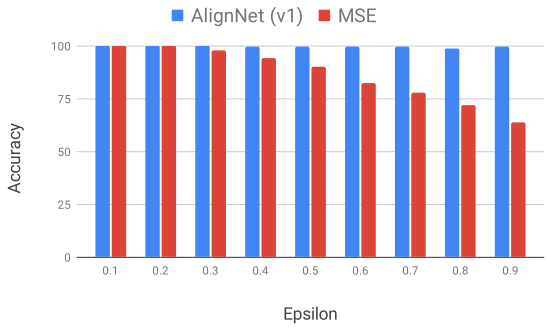


Figure 4.5: Classification accuracy on a simulated version of the structured transformation problem using the Symbolic Objects data for increasing values of ϵ. The AlignNet can learn to attend to unchanging features of an object to solve the alignment almost perfectly event at high levels of noise which degrade the baseline model's (MSE) performance.

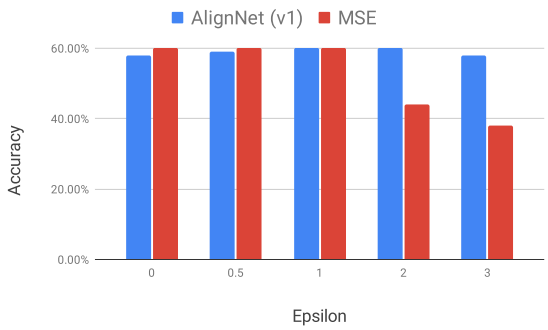


Figure 4.6: Classification accuracy on a simulated version of the structured transformation problem using the Learned Objects dataset for increasing values of ϵ. Note that due to multiple potential valid alignments over empty objects, the baseline accuracy without any noise is around 60. As seen in the Symbolic Objects dataset, the AlignNet can tolerate much higher levels of noise in comparison to the baseline model (MSE).

The Structured Transformation Problem

Using the Learned Objects dataset with our structured noise approach gave a systematic way of testing the AlignNet with learned object representations. However, our simulation of the structured transformation problem was an oversimplification. Ideally, we could use realistic object representations *and* realistic object transformations. To do this, we leveraged the fact that the Learned Objects dataset was sequential. Whereas previously we used $S = \text{Shuffle}(U)$, we now use $S = Shuffle(U_{t+\delta})$. For $\delta = 1$ this corresponds to trying to align objects in adjacent frames. As we increase δ, objects undergo more and more transformations subject to the structure of the physical dynamics in the scene from which they were drawn. Thus, this is a naturalistic domain to evaluate the AlignNet's ability to deal with structured transformations. However, with the dataset we had collected, there was no way to engage with the naturalistic structured transformation problem without also engaging with the effective cardinality change problem. As already mentioned, this incarnation of the AlignNet is not structured in a way that allows it to handle with the effective cardinality change problem. As such the absolute classification accuracies are not meaningful. Instead,

we report the AlignNet's improvement in classification accuracy over the baseline models.

In Figure 4.7, we show the AlignNet's classification accuracy (green) improvements over the baseline (pink) for $\delta \in [1, 2, 4, 8]$. Increased delta values are analogous to increases in ϵ in the previous section: the larger the delta, the larger the perturbations (on average) to object properties. We found that the AlignNet provides modest but consistent improvements over the baseline model suggesting that it has learned which object properties are indicative of object identity and which properties can change. However, it is also clear the introduction of the ECC problem had a significant impact here.

4.2.4 AlignNet V1.5

The immediately preceding section demonstrated the utility of the AlignNet to deal with the structured transformation problem. It also gave a clear example where the model was hindered by the ECC problem. To motivate a possible solution to the ECC problem, let's walk through a hypothetical example. Consider a case where U_t contains two vectors representing a red ball, r_t, and an orange ball, o_t. Furthermore, let's assume that at the next timestep, the orange ball has moved out of view and therefore U_{t+1} consists of two vectors: one which represents the red ball, r_{t+1} and the other which represents an empty object, e_{t+1} and, therefore, invoking the ECC problem. The correct alignment is to match $r_t \mapsto r_{t+1}$ and $o_t \mapsto e_{t+1}$. However, let's consider what actually would happen with the current version of the AlignNet. Under the same conditions where we would expect the AlignNet to handle structured transformations, we would expect the affinity from r_t to be highest for r_{t+1}. If we look at the affinity for o_t, however, it would almost definitely have a higher affinity for r_{t+1} than e_{t+1}. They are both balls, have similar colors, and differences in positions could be attributed to the fact that object positions tend to change over time. All

things considered, the red ball in the second frame is a good match for the orange ball. The issue is that the red ball in the second frame is an even better match to the red ball in the first frame. This thought experiments suggests there needs to be some communication about how well objects match before arriving on a final alignment. Thus, we introduce an additional step of computation to allow for the "arbitration" of attention activations.

Architecture

Previously, the AlignNet had the following form:

$$U' = MLP_{shared}(U) \text{ and } S' = MLP_{shared}(S) \tag{4.6}$$

$$Q = MLP_{queries}(U') \text{ and } K = MLP_{keys}(S') \tag{4.7}$$

$$\hat{A} = RowNorm(Q^T K) \tag{4.8}$$

To add our arbitration step, we add a second attention step directly over the *attention matrix*. Once the affinities have been calculated from the contents of U and S, we do not actually need to refer to the contents. However, we do need to provide a mechanism by which the queries and keys can address each other. As such, each row of the attention matrix is concatenated with a slot index which will be available as an address in the newly added attention step. The slot index takes the form of a scalar ranging from $[-1, 1]$ with a step size of $\frac{1}{K}$. We apply attention over the concatenation

of the attention matrix and slot index.

$$U' = MLP_{shared}(U) \text{ and } S' = MLP_{shared}(S) \tag{4.9}$$

$$Q_0 = MLP_{queries}(U') \text{ and } K_0 = MLP_{keys}(S') \tag{4.10}$$

$$\hat{A}_0 = [RowNorm(Q_0^T K_0); SlotIndex] \tag{4.11}$$

$$Q_A = MLP_{A_Q}(A_0) \text{ and } K_A = MLP_{A_K}(A_0) \tag{4.12}$$

$$\hat{A} = RowNorm(Q_A^T K_A) \tag{4.13}$$

The Effective Cardinality Change Problem

We find that this extra arbitrating attention step is sufficient to solve, to some extent, the ECC problem. Figure 4.7 shows the behavior of the improved AlignNet as we increase the value of δ. As we saw with the previous dataset, due to uncertainty about how to align multiple empty slots we see a starting accuracy of 60%. From this reference point, we see that the improved version of the AlignNet drastically outperforms the baseline model and the original AlignNet model as δ increases. Besides quantitative improvements, visual inspection of alignment results confirmed the model was capable of successfully handling the effective cardinality change problem.

4.2.5 Conclusion

In this section, we described two fundamental issues in object tracking: the transformation problem (which is a general object tracking problem) and the effective cardinality change problem (which is specific to object tracking in models that can have an "empty object"). To investigate these issues, we designed alignment tasks that systematically introduced the individual problems and gradually increased the difficulty of the alignment task. To actually solve these issues, we introduced two versions of the AlignNet architecture and paired it with a self-supervised learning

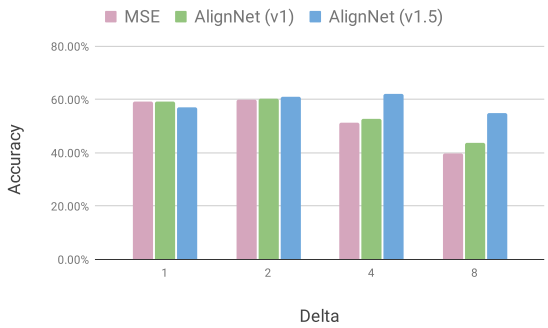


Figure 4.7: Classification accuracy for aligning across various temporal gaps (controlled by δ) in the Learned Objects dataset. We compare the baseline, no-learning approach "(MSE)" against two versions of the AlignNet. AlignNet (v1) is fundamentally incapably of handling the ECC problem, but still fairs better than the baseline as δ increases. Adding the second attention step in AlignNet (v1.5) substantially improves performance for the maximum value of $delta = 8$ where the ECC problem is more frequently triggered.

procedure wherein a model learns to align by shuffling objects to create an alignment task and training to unshuffle them. We found our model to perform accurate alignment even in cases where objects changed their appearances and were dropped or added. Future work should attempt to solve intertemporal alignment without the need for labelled data.

5

Conclusion

5.1 Summary of Thesis Work

The primary contributions of my thesis have been to investigate the conditions under which deep learning models learn intuitive physics and to subsequently push that capability. We ported the VoE paradigm from developmental psychology into a suite of synthetic probes which could be used to quantify knowledge of intuitive physics. We built several deep learning models with various inductive biases and demonstrated only a model with object inductive biases was capable of extrapolating concepts from intuitive physics based on generic experience. Finally, we built two models that provided a mechanism to learn stable (tracked) object decomposition which could eventually be slotted into the object-centric model to yield a comprehensive system for learning intuitive physics purely from visual experience.

In Chapter 2, we built a model with minimal inductive biases and assessed its ability to learn intuitive physics. To do so, we created a synthetic dataset comprising replications of classic paradigms in developmental psychology based on the VoE paradigm. To actually train the model, we built additional stimuli that contained variants of the events depicted in those classic paradigms. We found that this fairly

undifferentiated model – without any object-centric representations or computations – was able to learn physical concepts only in a narrow regime. Only when this model was exposed to the events seen in the same class as the probe trials was it capable to "pass" our VoE assessments on various physical concepts on the basis of generic experiences.

In Chapter 3, we built a model that adopted several inductive biases hypothesized in the developmental literature to guide infant knowledge of intuitive physics. Specifically, we built a model with a notion of object perception: it represented visual input as a set of objects. Furthermore, we endowed it with object tracking: the set of objects in the current image were linked to the corresponding objects in the previous image. Finally, we built an architecture that leveraged relational processing between objects to make per-object predictions. To evaluate this model, we revised our initial VoE dataset into the "Physical Concepts" dataset. This dataset improved over our previous efforts by using a splicing procedure to generate bias-free probes where the same images were present in both physically-possible and physically-impossible probes. This assured us that any VoE effects observed were due to violations of physical principles and not a result of simple changes in static image statistics. We found that these object-centric inductive biases allowed our model to learn intuitive physics from generic visual experience. Ablated models – which removed object-centric computation and representation but controlled for representational or parameter capacity – confirmed that without those inductive biases the model could not learn all the physical concepts tested. Finally, we analyzed the amount of visual experience needed to learn physical concepts and found it to be on the order of approximately 28 hours.

In Chapter 4, we paid down some of the explanatory debt incurred when endowing our PLATO model with object perception and tracking. By providing our model with ground truth object masks (for object perception and tracking), PLATO did not speak to how object perception or tracking may be learned. If we need objects to learn

intuitive physics, then an account of learning intuitive physics should address where objects come from in the first place. Our work with SlowMONet sought to give an account of how object perception (discovery) and tracking could inform each other to the benefit of both. We extended the MONet model to operate over video sequences and added a penalty for rapidly changing object representations – encouraging the model to keep objects in stable slot positions throughout a sequence. We found partial success with this method: we demonstrated it's ability to track identical objects purely based on motion, but were unable to get reliable object discovery in our intuitive physics dataset. Additionally, our work with an early version of the AlignNet isolated the problem of object tracking. The AlignNet outlined a successful approach to tracking a set of objects over time, which may undergo property changes or removals or additions to the set. Despite using partial supervision for some of the tasks, this work formed the basis for a newer version of the AlignNet (Creswell et al., 2020a) which can learn object tracking from unsupervised data. The combination of these approaches suggests that with additional engineering work we can build a system like PLATO which learns entirely unsupervised.

5.2 Future Work

5.2.1 End-to-end, Unsupervised Learning of Intuitive Physics

PLATO demonstrated that we could learn intuitive physics given tracked object masks in the service of object perception and tracking. Our work on SlowMONet, the initial version of AlignNet, and unsupervised AlignNet (work done by our colleagues) suggest it should be possible to train a system which can learn object perception and tracking entirely from scratch. Although we expect it to be a non-trivial endeavor, combining a system that learns object perception and tracking in an unsupervised fashion with our model which learns object dynamics in an unsupervised fashion seems like a natu-

ral avenue for future work. In doing so, such a model would provide a more complete account of how intuitive physics could be learned purely from experience and the amount of experience needed to learn various physical concepts. Of course, such a model would still have some explanatory debt: where does the object-centric architecture come from? I suspect that maybe even that question could be meaningfully explored with some advances in neural architecture search (Zoph and Le, 2016).

5.2.2 Expanded Physical Concepts

When we refined our initial VoE dataset to form the Physical Concepts dataset, we were constrained by a few factors. First, we had to focus on a small number of physical concepts as a proof of concept. Additionally, we had to choose physical concepts that were relevant in the developmental literature *and* were amenable to the splicing procedure which yielded bias-free probes. However, there are many more physical concepts in the developmental literature which are relevant to understanding intuitive physics such as the principle of support (unsupported objects fall, partially-supported objects sometimes fall), containment (objects move with their containers), etc. Beyond just increasing to more physical concepts involving rigid objects, future work should also target different material types. As early as 5 months of age, infants have been shown to have different expectations for solids and liquids (Hespos et al., 2009)

5.2.3 Real-world Data For Deep Neural Networks

An obvious limitation of our dataset is its synthetic nature. This introduces a superficial visual discrepancy between the objects in our dataset and real-world visual statistics. More importantly, we generated the physical events in our training dataset according to some heuristics which we believed would give the model ample opportunity to learn intuitive physics. This introduced a gap between the distribution of

physical events in our dataset and that which a young infant would experience. The work of (Smith and Slone, 2017) suggests this is not merely a hypothetical discrepancy: the visual experience of infants is highly skewed in early infancy in various ways throughout development. For maximum insight on infant intuitive physics, it would be appropriate to train models on real-world data which capture both the visual appearances and distribution of events experienced by actual infants. Recently, researchers have actually started collecting datasets that will eventually enable this kind of research. Sullivan et al. (2020) introduce the SAYCam corpus which captured the experiences of infants between the ages of 6-32 months by mounting a head camera for approximately 2 hours a week. (Clerkin et al., 2017b) collected a similar dataset using head-mounted cameras on infants between the ages of 8 and 11 months. The combination of this new class of datasets with likely improvements to object discovery models would be an insightful tool for investigating the development of intuitive physics in infants.

5.2.4 Synthetic Data For Real-world Babies

PLATO broadly demonstrated that object-centric inductive biases can enable a system to learn intuitive physics. We compared this to developmental accounts of intuitive physics which posited the necessity of objects. We caveated this comparison by highlighting it was based on several levels of abstractions. Amongst other things (e.g. learning substrate and mechanisms), we abstracted behavioral probes of intuitive physical knowledge. Our model used synthetic, procedurally-generated datasets. Meanwhile, infants were assessed using a specific set of displays and experiments instantiated in the real world using stereoscopic vision. While current DNN models are not sophisticated enough to give meaningful behavior on real-world data, the other direction is viable: infants can handle synthetic data. Thus, a prospect for fruitful research would be to present the datasets developed in this thesis to infants. This

would allow for much richer and detailed comparisons between our model (or subsequently developed models) and infant behavior. For example, comparing trial-by-trial surprise across DNNs and infants. Furthermore, if infants failed to show VoE effects on our dataset, that would serve as useful feedback to improve our dataset to better reflect intuitive physics knowledge.

5.2.5 Manipulating and Measuring the Statistics of Visual Experience

Whereas the Core Knowledge theory hypothesizes a set of constraints that guide infant learning in response to experiences, empiricist accounts of the development of infant knowledge appeal purely to the statistics of experiences. Such empiricist accounts, in contrast to constructivist accounts, downplay the *interaction* between experience and internal ideas (e.g. Piaget's formulation of the child as a scientist (Gratch, 2013)). Instead, knowledge is accrued via a gradual acquisition of the statistical regularities present in the environment. This implies that changing the distribution of experiences (e.g. adding a disproportionate number of containment events) should lead to the acquisition of different physical concepts (e.g. learning the principles of containment sooner than expected). An implication of the Core Knowledge hypothesis is that the order of acquisition of physical concepts is not random. Object solidity and permanence pave the way for learning subsequent physical principles. Future experiments which systematically vary the distribution of visual experience and measure the ordering of learned physical concepts could clarify whether the core physical principles in "Core Knowledge" are indeed a prerequisite for learning subsequent physical concepts or if so-called core concepts emerge early simply due to the "typical" distribution of visual experiences accrued by infants.

The experiment recommended above with DNNs would benefit from a complimentary set of developmental experiments to establish what "typical" visual experiences

look like in infants *across the world*. We laud recent work that has started to log infant visual experience using head-cam mounted data (Clerkin et al., 2017b; Borjon et al., 2018; Sullivan et al., 2020). We note, however, that all of these datasets were collected from infants in the United States of America. Future experiments should expand this enterprise to a wider range of locations and cultures. To make claims that some early knowledge is independent (the "Core Knowledge" hypothesis), partially dependent or largely dependent (the Constructivist or Empiricist views) on visual experience, scientists must ensure their data captures the true diversity of visual experience in infants across the world.

In summary, the proposed future work in this section suggests to study the effect of visual experiences on intuitive physical knowledge in two complimentary ways. First, using DNNs and synthetic data, we can *manipulate* the distribution of visual experiences and observed the emergence of various physical concepts. This line of research is directly enabled by the work in this thesis. Second, observe the true variety of visual statistics across various cultures and measure any variations in the order of physical concept acquisition. Such a multi-pronged approach would invariably shed light on "the origins of knowledge."

Bibliography

Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. In *Advances in neural information processing systems*, pages 5074–5082, 2016.

Andrèa Aguiar and Renée Baillargeon. 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive Psychology*, 39(2):116 – 157, 1999.

Renee Baillargeon. Object permanence in 31/2-and 41/2-month-old infants. *Developmental psychology*, 23(5):655, 1987.

Renée Baillargeon. The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell handbook of childhood cognitive development*, 1(46-83):1, 2002.

Renee Baillargeon. Infants' reasoning about hidden objects: evidence for event-general and event-specific expectations. *Developmental science*, 7(4):391–414, 2004.

Renée Baillargeon. Innate ideas revisited: For a principle of persistence in infants' physical reasoning. *Perspectives on Psychological Science*, 3(1):2–13, 2008.

Renée Baillargeon and Susan Carey. Core cognition and beyond the acquisition of physical and numerical knowledge. In Sabina Pauen, editor, *Early Childhood Development and Later Outcome*, chapter 3, pages 33–65. Cambridge University Press, 2012a.

Renée Baillargeon and Susan Carey. Core cognition and beyond: The acquisition of physical and numerical knowledge. *Early childhood development and later outcome*, pages 33–65, 2012b.

Renée Baillargeon and Julie DeVos. Object permanence in young infants: Further evidence. *Child development*, 62(6):1227–1246, 1991.

Renée Baillargeon, Elizabeth S. Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985a.

Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985b.

Renée Baillargeon, Amy Needham, and Julie DeVos. The development of young infants' intuitions about support. *Early development and parenting*, 1(2):69–78, 1992.

Pierre Baldi and Laurent Itti. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666, 2010.

W. A. Ball. The perception of causality in the infant. In *Society for Research in Child Development*, 1973.

Victor Bapst, Alvaro Sanchez-Gonzalez, Carl Doersch, Kimberly L Stachenfeld, Pushmeet Kohli, Peter W Battaglia, and Jessica B Hamrick. Structured agents for physical construction. *arXiv preprint arXiv:1904.03177*, 2019.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Neural Information Processing Systems*, NIPS'16, 2016.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 2013.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Jeremy I Borjon, Sara E Schroer, Sven Bambach, Lauren K Slone, Drew H Abney, David J Crandall, and Linda B Smith. A view of their own: Capturing the egocentric view of infants and toddlers with head-mounted cameras. *Journal of visualized experiments: JoVE*, (140), 2018.

Matthew Botvinick, David G. T. Barrett, Peter Battaglia, Nando de Freitas, Darshan Kumaran, Joel Z Leibo, Timothy Lillicrap, Joseph Modayil, Shakir Mohamed, Neil C. Rabinowitz, Danilo J. Rezende, Adam Santoro, Tom Schaul, Christopher Summerfield, Greg Wayne, Theophane Weber, Daan Wierstra, Shane Legg, and Demis Hassabis. Building machines that learn and think for themselves. *Behavioral and Brain Sciences*, 40, 2017.

Thomas GR Bower. The object in the world of the infant. *Scientific American*, 225 (4):30–39, 1971.

J. Gavin Bremner, Alan M. Slater, and Scott P. Johnson. Perception of object persistence: The origins of object permanence in infancy. *Child Development Perspectives*, 9(1):7–13, 2015.

Christopher P. Burgess, Loïc Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew Botvinick, and Alexander Lerchner. Monet: Unsupervised

scene decomposition and representation. *CoRR*, abs/1901.11390, 2019. URL http://arxiv.org/abs/1901.11390.

Rosa Cao and Daniel Yamins. Explanatory models in neuroscience: Part 1 – taking mechanistic abstraction seriously, 2021.

Susan Carey. The origin of concepts. *Journal of Cognition and Development*, 1(1): 37–41, 2000.

Susan Carey and Elizabeth Spelke. Science and core knowledge. *Philosophy of science*, 63(4):515–533, 1996.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

Michael Chang, Tomer Ullman, Antonio Torralba, and Joshua Tenenbaum. A compositional object-based approach to learning physical dynamics. In *International Conference on Learning Representations*, 2017a.

Michael Chang, Tomer Ullman, Antonio Torralba, and Joshua Tenenbaum. A compositional object-based approach to learning physical dynamics. In *International Conference on Learning Representations*, 2017b.

Cinzia Chiandetti and Giorgio Vallortigara. Intuitive physical reasoning about occluded objects by inexperienced chicks. *Proceedings of the Royal Society B: Biological Sciences*, 278(1718):2621–2627, 2011.

Elizabeth M Clerkin, Elizabeth Hart, James M Rehg, Chen Yu, and Linda B Smith. Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B*, 372(1711):20160055, 2017a.

Elizabeth M Clerkin, Elizabeth Hart, James M Rehg, Chen Yu, and Linda B Smith. Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160055, 2017b.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Antonia Creswell, Kyriacos Nikiforou, Oriol Vinyals, Andre Saraiva, Rishabh Kabra, Loic Matthey, Chris Burgess, Malcolm Reynolds, Richard Tanburn, Marta Garnelo, et al. Alignnet: Unsupervised entity alignment. *arXiv preprint arXiv:2007.08973*, 2020a.

Antonia Creswell, Luis Piloto, David Barrett, Kyriacos Nikiforou, David Raposo, Marta Garnelo, Peter Battaglia, and Murray Shanahan. Alignnet: Self-supervised alignment module, 2020b. URL https://openreview.net/forum?id=H1gcw1HYPr.

Misha Denil, Pulkit Agrawal, Tejas D Kulkarni, Tom Erez, Peter Battaglia, and Nando de Freitas. Learning to perform physics experiments via deep reinforcement learning. In *International Conference on Learning Representations*, 2017.

Jeffrey L Elman, Elizabeth A Bates, Mark H Johnson, Annette Karmiloff-Smith, Kim Plunkett, and Domenico Parisi. *Rethinking innateness: A connectionist perspective on development*, volume 10. MIT press, 1998.

S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models, 2016.

Scott E Fahlman and Christian Lebiere. The cascade-correlation learning architecture. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 1990.

Robert L Fantz. Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, 146(3644):668–670, 1964.

C. Finn and S. Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793, 2017.

Mevlana Gemici, Chia-Chun Hung, Adam Santoro, Greg Wayne, Shakir Mohamed, Danilo J. Rezende, David Amos, and Timothy Lillicrap. Generative temporal models with memory. 2017. URL http://arxiv.org/abs/1702.04649.

James J Gibson. The ecological approach to visual perception. 1979.

Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In *ICLR*, 2020.

Gerald Gratch. chapter 3: Recent studies based on piaget's view of object concept development. *Infant perception: From sensation to cognition: Perception of space, speech, and sound*, 2:51, 2013.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loïc Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. *CoRR*, abs/1903.00450, 2019. URL http://arxiv.org/abs/1903.00450.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Susan J Hespos and Renée Baillargeon. Infants' knowledge about occlusion and containment events: A surprising discrepancy. *Psychological Science*, 12(2):141–147, 2001a.

Susan J Hespos and Renée Baillargeon. Reasoning about containment events in very young infants. *Cognition*, 78(3):207 – 245, 2001b.

Susan J. Hespos and Renée Baillargeon. Infants' knowledge about occlusion and containment events: A surprising discrepancy. *Psychological Science*, 12(2):141–147, 2001c.

Susan J Hespos and Kristy vanMarle. Physics for infants: Characterizing the origins of knowledge about objects, substances, and number. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(1), 2012.

Susan J Hespos, Alissa L Ferry, and Lance J Rips. Five-month-old infants have different expectations for solids and liquids. *Psychological Science*, 20(5):603–611, 2009.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

S Hochreiter and J Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–80, 1997a.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997b.

Bärbel Inhelder and Jean Piaget. *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*, volume 22. Psychology Press, 1958.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

Mark Johnson. *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press, 2013.

Zsuzsa Káldy and Natasha Sigala. The neural mechanisms of object working memory: what is where in the infant brain? *Neuroscience & Biobehavioral Reviews*, 28(2): 113–121, 2004.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. URL http://arxiv.org/abs/1705.06950.

Frank C Keil. *Semantic and conceptual development: an ontological perspective*. Harvard University Press, 1979.

Melissa M Kibbe and Alan M Leslie. What's the object of object working memory in infancy? unraveling 'what'and 'how many'. *Cognitive Psychology*, 66(4):380–404, 2013.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. *arXiv preprint arXiv:1911.12247*, 2019.

David Klahr and Robert S Siegler. The representation of children's knowledge. In *Advances in child development and behavior*, volume 12, pages 61–116. Elsevier, 1978.

Adam R. Kosiorek, Hyunjik Kim, Ingmar Posner, and Yee Whye Teh. Sequential attend, infer, repeat: Generative modelling of moving objects, 2018.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *International Conference on International Conference on Machine Learning*, ICML 2016, 2016.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition, 2020.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *arXiv preprint arXiv:2006.15055*, 2020.

Jean M Mandler. How to build a baby: Ii. conceptual primitives. *Psychological review*, 99(4):587, 1992.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. 1982. ISBN 0716715678.

James L McClelland. Parallel distributed processing: Implications for cognition and development. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA ARTIFICIAL INTELLIGENCE AND PSYCHOLOGY, 1988.

Albert Michotte. *The perception of causality*. Basic Books, 1963.

Tom M Mitchell et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37): 870–877, 1997.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Yuko Munakata. Challenges to the violation-of-expectation paradigm: Throwing the conceptual baby out with the perceptual processing bathwater? *Infancy*, 1(4): 471–477, 2000.

Amy Needham and Renee Baillargeon. Intuitions about support in 4.5-month-old infants. *Cognition*, 47(2):121–148, 1993.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W Battaglia. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409*, 2020.

Jean Piaget and Margaret Trans Cook. The origins of intelligence in children. 1952.

Jean Piaget and Margaret Trans Cook. The construction of reality in the child. 1954.

Luis Piloto, Ari Weinstein, Dhruva TB, Arun Ahuja, Mehdi Mirza, Greg Wayne, David Amos, Chia-Chun Hung, and Matthew Botvinick. Probing physics knowledge using tools from developmental psychology. *arxiv*, abs/1804.01128, 2018. URL http://arxiv.org/abs/1804.01128.

Luis Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Learning intuitive physics: A deep-learning model inspired by developmental psychology. *submitted*, 2021.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. 2018. URL https://coml.lscp.ens.fr/intphys/.

Susan M Rivera, Ann Wakeley, and Jonas Langer. The drawbridge phenomenon: Representational reasoning or perceptual preference? *Developmental psychology*, 35(2):427, 1999.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, pages 8459–8468. PMLR, 2020.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *Neural Information Processing Systems*. 2017a.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017b.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Thomas R Shultz, Denis Mareschal, and William C Schmidt. Modeling cognitive development on balance scale phenomena. *Machine learning*, 16(1):57–86, 1994.

Pranav Shyam, Shubham Gupta, and Ambedkar Dukkipati. Attentive recurrent comparators. In *International Conference on Machine Learning*, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Brian Cantwell Smith. *The promise of artificial intelligence: reckoning and judgment*. Mit Press, 2019.

Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbaum, and Tomer Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems*, pages 8983–8993, 2019.

Linda B Smith and Lauren K Slone. A developmental approach to machine learning? *Frontiers in psychology*, 8:2124, 2017.

Elizabeth Spelke. Initial knowledge: six suggestions. *Cognition*, 50(1):431 – 445, 1994.

Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992.

Elizabeth S Spelke, Gary Katz, Susan E Purcell, Sheryl M Ehrlich, and Karen Breinlinger. Early knowledge of object motion: Continuity and inertia. *Cognition*, 51 (2):131–176, 1994a.

Elizabeth S Spelke, Gary Katz, Susan E Purcell, Sheryl M Ehrlich, and Karen Breinlinger. Early knowledge of object motion: Continuity and inertia. *Cognition*, 51 (2):131–176, 1994b.

Elizabeth S Spelke, Roberta Kestenbaum, Daniel J Simons, and Debra Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, 13(2):113–142, 1995.

Jess Sullivan, Michelle Mei, Amy Perfors, Erica H Wojcik, and Michael C Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. 2020.

Ernő Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B. Tenenbaum, and Luca L. Bonatti. Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 332(6033), 2011.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature analysis. *Neural computation*, 19(4):1022–1038, 2007.

Tomer D. Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B. Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9):649 – 665, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*. 2017.

Rishi Veerapaneni, John D Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua Tenenbaum, and Sergey Levine. Entity abstraction in visual model-based learning. In *Conference on Robot Learning*, pages 1439–1456, 2020.

Su-hua Wang and Renée Baillargeon. Can infants be "taught" to attend to a new physical variable in an event category? the case of height in covering events. *Cognitive Psychology*, 56(4):284–326, 2008.

Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *Advances in neural information processing systems*, pages 4539–4547, 2017.

Nick Watters, Loic Matthey, Chris P Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for disentangled representations in vaes. 2019.

Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Neural Information Processing Systems*. 2015.

Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In *Neural Information Processing Systems*. 2017.

Karen Wynn. Addition and subtraction by human infants. *Nature*, 358(6389):749–750, 1992.

Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. Relational deep reinforcement learning. *arXiv preprint arXiv:1806.01830*, 2018.

Bo Zheng, Yibiao Zhao, Joey Yu, Katsushi Ikeuchi, and Song-Chun Zhu. Scene understanding by reasoning stability and safety. *International Journal of Computer Vision*, 112(2):221–238, 2015.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.