A Survey on Machine Learning Approaches for Modelling Intuitive Physics

Jiafei Duan[1*], **Arijit Dasgupta**[2*], **Jason Fischer**[3] and **Cheston Tan**[1,4]
[1]Institute for Infocomm Research, A*STAR
[2]Department of Mechanical Engineering, National University of Singapore
[3]Department of Psychological and Brain Sciences, Johns Hopkins University
[4]Centre for Frontier AI Research, A*STAR
{duan_jiafei, cheston-tan}@i2r.a-star.edu.sg, arijit.dasgupta@u.nus.edu, jason.fischer@jhu.edu

Abstract

Research in cognitive science has provided extensive evidence of human cognitive ability in performing physical reasoning of objects from noisy perceptual inputs. Such a cognitive ability is commonly known as intuitive physics. With advancements in deep learning, there is an increasing interest in building intelligent systems that are capable of performing physical reasoning from a given scene for the purpose of building better AI systems. As a result, many contemporary approaches in modelling intuitive physics for machine cognition have been inspired by literature from cognitive science. Despite the wide range of work in physical reasoning for machine cognition, there is a scarcity of reviews that organize and group these deep learning approaches. Especially at the intersection of intuitive physics and artificial intelligence, there is a need to make sense of the diverse range of ideas and approaches. Therefore, this paper presents a comprehensive survey of recent advances and techniques in intuitive physics-inspired deep learning approaches for physical reasoning. The survey will first categorize existing deep learning approaches into three facets of physical reasoning before organizing them into three general technical approaches and propose six categorical tasks of the field. Finally, we highlight the challenges of the current field and present some future research directions.

1 Introduction

Humans have demonstrated the innate ability to approximate predictions of their surrounding interactions and physical environment even without any formal education in physics [McCloskey *et al.*, 1983a] as shown in the examples from Figure 1. In fact, research in developmental psychology shows that infants as young as two and half months can understand fundamental physics [Carey, 2000; Baillargeon, 2004] and by three months old, they can detect violations of physical principles of persistence, continuity and solidity [Leslie, 1984]. They achieve this separately from acquisition of semantic knowledge, language, and sensorimotor skills. This cognitive capability of humans is commonly termed **intuitive physics**, and widely used by researchers in multiple disciplines like cognitive science, neuroscience and computer science.

For several decades, the cognitive science perspectives on intuitive physics have been shaped by the question of how humans acquire and deploy this cognitive capability. Notwithstanding various debates on how intuitive physics works, several conventional and widely plausible ideas have been proposed by cognitive scientists. They include heuristic or rule-based models [Gilden and Proffitt, 1994; Runeson *et al.*, 2000; Sanborn *et al.*, 2013], probabilistic mental simulation [Hegarty, 2004; Bates *et al.*, 2015], and the cognitive intuitive physics engine (IPE) [Battaglia *et al.*, 2013; Ullman *et al.*, 2017], each with its own merits. With recent advancements of computing technology and motivations to create machines that can learn and think like humans [Lake *et al.*, 2017], the modern field of intuitive physics has been reinvigorated by new techniques in artificial intelligence (AI).

This paper aims to provide a comprehensive survey of intuitive physics for machine cognition, covering the recent advancements in computer vision, deep learning, and AI deployed to model human-level intuitive physics capability for various physical reasoning tasks. This survey will first examine intuitive physics in machine cognition by categorizing existing works into three facets of physical reasoning, namely *prediction, inference and causal reasoning*. Then, the paper will categorize the space of physical reasoning tasks into six categorical tasks, and further review them via their three general technical approaches (**inverse rendering, inverse physics, and inverse dynamics**) as shown in Table 1. Lastly, the paper will then conclude with some of the open problems, challenges and future trends in intuitive physics for machine cognition.

There is a larger body of work that could constitute as intuitive physics for machine cognition. However, this survey is scoped specifically to the facets, approaches and tasks defined in later sections. For instance, physical reasoning with 3D and 2.5D data representations [Zheng *et al.*, 2013; Jia *et al.*, 2014; Du *et al.*, 2018], or on the non-rigid body (e.g. fluid, particle, and soft-body) [Li *et al.*, 2018; Mrowca *et al.*, 2018; Li *et al.*, 2020b], and learning physics via action-task planning [Song and Boularias, 2018; Bakhtin *et al.*, 2019; Xu *et al.*, 2020] will not be covered in this paper.

*Equal Contributions

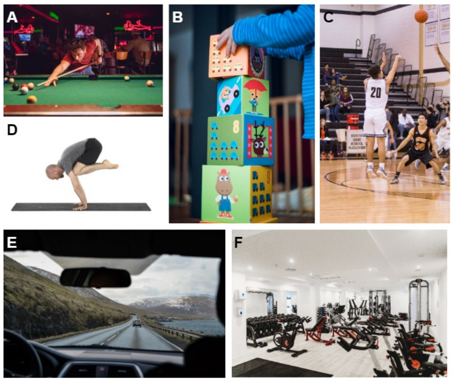

Figure 1: Examples of everyday scenes that requires us to employ intuitive physics. (A) Predicting the trajectories of the billiard balls. (B) Balancing the stacking blocks. (C) Shooting the basketball with a parabola trajectory. (D) Balancing one's body during yoga. (E) Estimating the velocity of the car travelling ahead. (F) A gym with equipment that exhibits various degree of physical properties.

2 Background

2.1 Motivation

A primary purpose of intuitive physics is to allow us to plan effective actions on the world. We have a goal in mind for a physical outcome we want to achieve, and intuitive physics allows us to assess the consequences of possible actions in order to select those that will achieve our goals. However, intuitive physics can be approximate, but it shouldn't be fundamentally wrong. But yet, humans tend to develop various misconceptions [Caramazza *et al.*, 1981; McCloskey *et al.*, 1983b] in their physical judgement (e.g., the Aristotelian prediction). The main goal of intuitive physics in human cognition is our ability to rely on intuition and build upon the interactions with our surroundings and make adequate physical reasoning of the observed events. Prior work in cognitive science literature emphasizes the importance of intuitive physics as forms of high-level reasoning capabilities rooted in developing intelligence systems. Therefore, intuitive physics understanding is vital for AI to have a general understanding of physical scenes and ensuring the safety of embodied AI [Duan *et al.*, 2021b] systems deployed into the real-world [Duchaine and Gosselin, 2009; Zheng *et al.*, 2015]. Furthermore, there is a diverse amount of work in intuitive physics for machine cognition, however with a scarcity of comprehensive survey papers [Kubricht *et al.*, 2017] on the field from a machine learning perspective.

2.2 Survey Organization

The paper will look into current efforts in physical reasoning via a hierarchical structure by first categorizing them into three facets of physical reasoning: *prediction, inference* and *causal reasoning* [Smith *et al.*, in press]. Following that, the paper groups existing efforts for intuitive physics for machine cognition by three general technical approaches for physical reasoning tasks: **inverse rendering**, which uses a single image to extrapolate and learn useful features, **inverse physics**, which uses latent physical representation (e.g., object relation graph, physical properties, and others) to perform physical reasoning, and lastly **inverse dynamics**, which uses sequential roll-out frames as a representation of object dynamics in physical events for learning. The work under these three general approaches is further categorised based on six physical reasoning tasks widely used to evaluate intuitive physics for machine cognition. The six physical reasoning tasks are Physical Interaction Outcome prediction (PIO), Physical Trajectories/Dynamics prediction (PTD), Physical Properties Inference (PPI), Visual State Generation (VSG), Violation-of-Expectation detection (VoE), and other intuitive physics-inspired AI tasks (Others). The details of these six physical reasoning tasks are illustrated in Figure 2. The paper evaluates all existing work by their approaches, physical reasoning tasks and evaluation metrics in Section 3 and Table 1. Finally, the paper will discusses the current challenges of the field and proposes some open questions in Section 4.4.

With a few exceptions (e.g. generative task, counterfactual prediction, or predicting physical equations), the majority of physical reasoning tasks focus on the machine learning task of classification. This is due to the broad nature of humans' ability to convey the output of our intuitive physics model via a form of classifying the various possibilities of any physical interaction and thus provide reasoning through classification.

3 Facets of Physical Reasoning

3.1 Prediction

The goal of *prediction* in physical reasoning, from a cognitive standpoint, is to deploy a forward intuitive physics model for physical interactions before querying on the simulated outcomes and making judgments about the possible future states of the interactions. The goal of *prediction* for machine cognition focuses on a simple notion, which is taking in visual inputs and performing physical reasoning of the queried scenarios. All work cited in this section tackled this facet via deep learning. The deep neural networks (DNN) implemented generally mapped the visual inputs to the various physical prediction outputs (e.g., the outcome of the physical interactions, physical trajectories, inferred physical properties of objects and generating future frames).

PIO, **PTD**, **PPI**, and **VSG** predictions are the common physical reasoning tasks that fall under the facets of *prediction*. Most of the work under *prediction* employ either an inverse rendering or inverse dynamics approach, with one exception [Mottaghi *et al.*, 2016b]. They focused on using conventional convolutional neural network (CNN) frameworks such as InceptionNet [Szegedy *et al.*, 2015], AlexNet [Iandola *et al.*, 2016], and ResNet [He *et al.*, 2016] as their backbone for learning to map the input pixels into low-level features that are later used to predict the various physical reasoning tasks. As a result, many of these works came at the beginning of deep learning era. Consequently, all of these works [Mottaghi *et al.*, 2016a; Li *et al.*, 2016; Lerer *et al.*, 2016; Mottaghi *et al.*, 2016b; Groth *et al.*, 2018; Janner *et al.*, 2018; Watters *et al.*, 2017; Wu *et al.*, 2017; Ehrhardt *et al.*, 2019b; Duan *et al.*, 2021a] fall under the facet of *prediction*. Within the facets of physical reasoning, the falling-tower test is one of the most common tests for evaluating a model's ability in

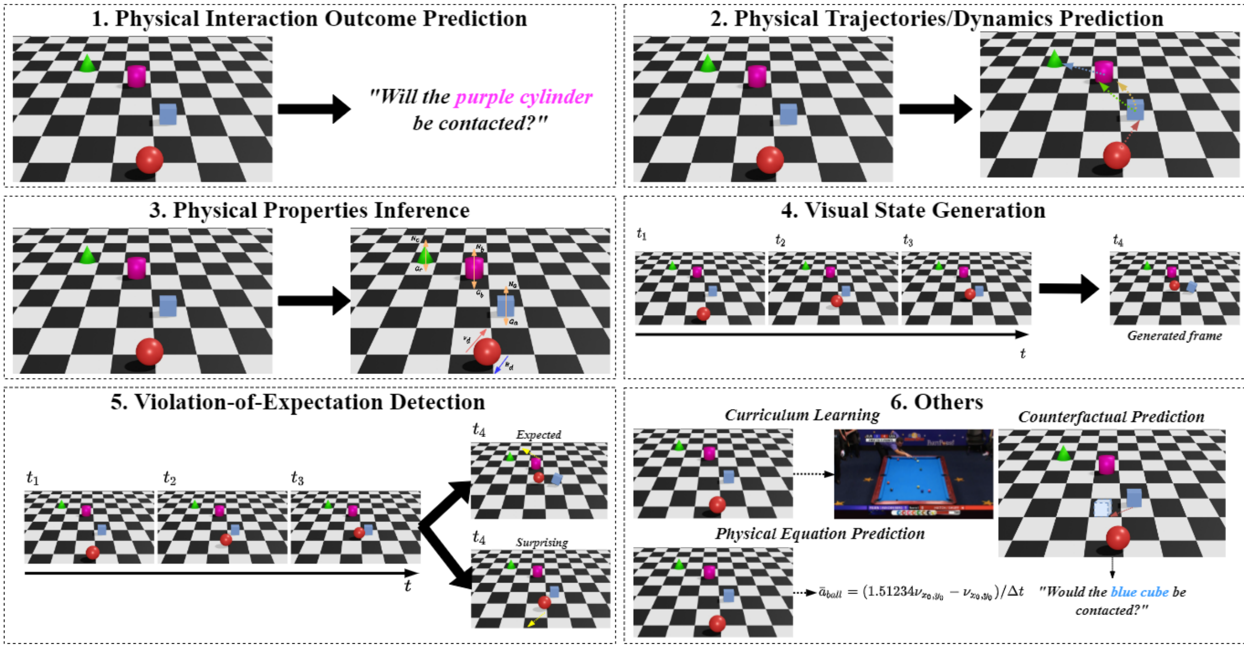

Figure 2: Summary of the six physical reasoning tasks. (1) **PIO**, to predict the different states or outcome of physical interactions (e.g., *if objects within the dynamic scene is stable, contained, or contacted*). (2) **PTD**, to predict the possible physical trajectories given only a few dynamics scenes. (3) **PPI**, to infer both the observed(e.g., *size, color, and shape*) and latent physical properties (e.g., *mass, friction, velocity, and displacement*) of the objects within the dynamic scenes. (4) **VSG**, to generate the unseen future frames of a long roll-out sequence given only the initial few dynamics frames. (5) **VoE**, to classify if there is any violation-of-expectation in a given dynamics scenes. (6) **Others**, other intuitive physics-inspired AI tasks such as curriculum learning, counterfactual prediction, physical equations prediction etc.

physical reasoning. Hence, all of these works [Li *et al.*, 2016; Mottaghi *et al.*, 2016b; Lerer *et al.*, 2016; Groth *et al.*, 2018; Janner *et al.*, 2018; Duan *et al.*, 2021a] focus on predicting the outcome of stability for the given dynamics scenarios. Besides just predicting the outcome of stability, Lerer *et al.* [2016] even predicts the trajectories of the stacking blocks via segmentation masks, while Groth *et al.* [2018] predicts the point of instability through generating a heat-map. Duan *et al.* [2021a] predicts the outcome of physical interaction by mimic the "noisy" framework in human physical prediction with generative models, and using attention to focus on salient moments of physical interactions.

Another common test for physical reasoning is for the model to make judgments of the physical properties for object interactions in Newtonian motion (e.g. Spring, Gravity, Billiards, Magnetic billiards, drift). As a result of this test for intuitive physics, works such as Mottaghi *et al.* [2016a] and Watters *et al.* [2017] generate a set of synthetic 2D video dataset using a physics engine for the models' training. Their models would focus on first learning to infer the physical properties from a given visual input using DNN and use the derived physical properties to reason and generate the potential physical trajectories or dynamics of the objects within the scene. The work in Watters *et al.* [2017] even employs a physics engine to reverse the learned latent state back into generated frames forward in time. On the other hand, works such as Wu *et al.* [2017] and Ehrhardt *et al.* [2019b] use this test to evaluate their model's intuitive physics ability; however, they train their models using real-world video datasets collected on these various physical interactions. Wu *et al.* [2017] would infer object properties from visual inputs and, using a game engine, simulate these obtained physical properties back into an image. In contrast, Ehrhardt *et al.* [2019b] uses a meta-learning approach to learn from the past dynamics scenes and uses that to optimise the learning process of predicting new trajectories. While inferring physical properties is a step of their approaches, we consider these works to fall under the facet of prediction as their main goal in the tasks is predictive in nature.

3.2 Inference

The aim of inference in physical reasoning is to inductively represent the physical properties of a physical interaction. Consequently, most of the papers listed under the facet of *inference* tackle the physical reasoning task of **PPI** (with the exception of Wu *et al.* [2016] and Huang *et al.* [2018]). A common theme of papers under this facet is that they often have an additional component of a prediction task. Although one may construe such papers to also fall under the facet of *prediction*, we label these papers under *inference* as they form the best representative work of inference in intuitive physics and the inference stage forms the main and crucial back-end step of the models described in them. When defining physical properties, researchers generally consider two components.

The first component refers to extrinsic/observable physical properties (e.g. speed, size, shape, position), which are generally directly determined with little difficulty by humans. For instance, Huang *et al.* [2018] employs a two-stage Faster-

Facets of Physical Reasoning	Method/Category	Publication	Year	Physical Reasoning Tasks	Evaluation Metrics
Prediction	Inverse Rendering	[Mottaghi *et al.*, 2016a]	2016	PPI	F_1, MHD
		[Li *et al.*, 2016]	2016	PPI	$Acc_{\%}$, R^2
		[Lerer *et al.*, 2016]	2016	PIO, PTD	$Acc_{\%}$, IoU, Log_L
		[Groth *et al.*, 2018]	2018	PIO, PPI	$Acc_{\%}$
		[Janner *et al.*, 2018]	2018	VSG	$Acc_{\%}$, MSE
	Inverse Physics	[Mottaghi *et al.*, 2016b]	2016	PTD	$Acc_{\%}$
		[Ehrhardt *et al.*, 2019b]	2019	PTD	MSE, $Error$
	Inverse Dynamics	[Watters *et al.*, 2017]	2017	PTD, VSG	IN_L, *MEPE*
		[Wu *et al.*, 2017]	2017	PPI	MSE, MAE
		[Duan *et al.*, 2021a]	2021	PIO, VSG	$Acc_{\%}$, *PSNR*
Inference	Inverse Rendering	[Wu *et al.*, 2015]	2015	PIO, PPI	$Acc_{\%}$, MSE
		[Wu *et al.*, 2016]	2016	PIO, PTD	MSE, R^2
	Inverse Physics	[Battaglia *et al.*, 2016]	2016	PPI, PTD	MSE
		[Chang *et al.*, 2016]	2016	PPI	MSE, $Acc_{\%}$
		[Huang *et al.*, 2018]	2018	Others	R^2, *MEPE*
		[Xu *et al.*, 2021]	2021	PPI	IoU, MSE
	Inverse Dynamics	[Fragkiadaki *et al.*, 2016]	2016	PPI, PTD	$Error$, $Acc_{\%}$
		[Ye *et al.*, 2018]	2018	PIO, PPI	$Error$
		[Zheng *et al.*, 2018]	2018	PPI, VSG	R^2, *MEPE*
		[de Avila Belbute-Peres *et al.*, 2018]	2018	PPI, VSG	MSE
		[Jaques *et al.*, 2019]	2019	PPI, VSG	SSI
		[Kandukuri *et al.*, 2022]	2022	PPI	$Error$
Causal Reasoning	Inverse Rendering	[Piloto *et al.*, 2018]	2018	VoE	*KD*
		[Smith *et al.*, 2019]	2019	VoE	$Acc_{\%}$, RMSE
	Inverse Physics	[Baradel *et al.*, 2020]	2020	VSG, Others	MSE
		[Li *et al.*, 2020a]	2020	VSG, Others	MRR, F_1
		[Dasgupta *et al.*, 2021]	2021	VoE	$Acc_{\%}$
	Inverse Dynamics	[Riochet *et al.*, 2018]	2018	VoE	1-AUC
		[Ates *et al.*, 2020]	2020	PIO, Others	$Acc_{\%}$
		[Yi *et al.*, 2020]	2020	PIO, Others	$Acc_{\%}$

Table 1: Summary of the work for intuitive physics for machine cognition. Physical Reasoning Tasks: Physical Interaction Outcome prediction (PIO), Physical Trajectories/Dynamics prediction (PTD), Physical Properties Inference (PPI), Visual State Generation (VSG), Violation-of-Expectation detection (VoE), other intuitive physics-inspired AI tasks (Others). The evaluation metric: F1 score (F_1), modified hausdorff distance (MHD), prediction accuracy ($Acc_{\%}$), intersection over union (IoU), log likelihood (Log_L), mean squared error (MSE), prediction variable error ($Error$), mean euclidean prediction error (*MEPE*), inverse normalized loss (IN_L), coefficient of determination (R^2), structural similarity index (SSI), KL divergence (KD), mean reciprocal rank (MRR) and error rate (1-AUC).

RCNN model to infer the position and velocity of physical scenarios as a necessary step to learn governing physics equations via symbolic regression. However, the inference of physical properties revealed by surface textures are far easier to estimate than latent properties (e.g. mass, density, friction, coefficient of restitution, spring constant) that must be inferred through the observation of physical interactions. These latent properties cannot be directly observed and often requires humans to observe the outcome of an interaction before refining their judgements of such properties. This added challenge explains why the remainder of papers under this facet attempt to infer latent properties.

The majority of *inference* papers attempt to infer latent properties from physical interactions of bouncing balls, collisions, spring oscillation and an object moving on a ramp or level surface. This is because the outcome of these interactions are sensitive to latent properties like mass, coefficient of restitution (affecting the speed after a collision), friction (affecting an object's ability to slide on a surface) and spring constant (affecting an object-spring system oscillation). In Wu *et al.* [2015], a generative 3D physics engine coupled with an object tracking algorithm was used to infer the mass, position, shape and friction of an object sliding down a ramp. A markov chain monte-carlo approach was used to infer the likelihood of these physical properties with a real world dataset. Inspired by this 'analysis-by-synthesis' approach, Jaques *et al.* [2019] implemented an unsupervised deep learning encoder-decoder inverse graphics framework to infer the spring constant, gravity and mass of a 2D synthetic ball-spring system and 3-body gravitational system. Contrary to this inverse rendering approach, Wu *et al.* [2016] learns directly from visual inputs by first inferring observable physical properties and using them as supervision to infer latent physical properties. They also contributed a real-world dataset, Physics 101, which challenges models to infer these physical properties in various physical scenarios (e.g. object on a ramp, object floating/sinking in liquid).

One limitation thus far is that these approaches do not generalize to any number of objects. Hence, Battaglia *et al.* [2016] introduced the interaction network, a learnable graph model which can reason about the interactions with any number of objects with given object and relation based representations. This network allowed for the inference of the potential energy in 2D synthetic N-body collision interactions. The interaction network was adopted by Zheng *et al.* [2018] as a unit for a recurrent neural network (RNN) to infer the coefficient of restitution and mass of elastic and inelastic collisions. The interaction network was also adopted by Chang *et al.* [2016] who combined it with pairwise factorization, context selection and function composition to form the Neural Physics Engine (NPE), which can infer discrete mass values of a 2D synthetic dataset of bouncing balls.

The need to disentangle the visual properties of an object from its dynamics for the sake of generalizability was highlighted by Chang *et al.* [2016] as it is possible for objects to differ visually and have the same dynamics. Another form disentanglement was deemed necessary by Ye *et al.* [2018], who underscored that it is not possible to determine the exact values of mass and friction together as they are highly dependent quantities. Therefore, in their encoder-decoder deep learning approach for their 3D (in 2D video) collision dataset, they represented the mass and friction in the latent space of the bottleneck layer and staggered the training such that the model trained to infer each quantity separately.

Another approach to the facet of *inference* is the use of a differentiable physics engine. de Avila Belbute-Peres *et al.* [2018] proposed a 2D differentiable physics engine that is defined with a linear complementarity problem (LCP). To test their physics engine, they illustrated its capability in inferring the unknown masses of bouncing billiard balls. The approach in de Avila Belbute-Peres *et al.* [2018] was capable of fast inference via an analytical solution and had higher sample efficiency, unlike the earlier mentioned data-driven approaches. Kandukuri *et al.* [2022] also used a differentiable physics engine in their model as the first step in inferring the mass and friction of various physical scenarios (block on a flat/inclined plate and block collision). Finally, Xu *et al.* [2021] also addressed the issue of sample efficiently by introducing the Bayesian Symbolic Physics (BSP) model, a probabilistic learning approach that can infer the mass and friction of N-body and bouncing ball interactions with significantly fewer (10×) samples.

3.3 Causal Reasoning

The facet of causal reasoning in the context of physical reasoning can be defined via two paradigms: counterfactual reasoning (**Others**) and **VoE**. The counterfactual reasoning paradigm is inspired by the theory that humans reason about causal events (considering events/objects A, B and C; does A cause B to lead to C?) by mentally modelling the scenario if the absence of the proposed causal object/event would still lead to the same outcome (if A did not exist, would B still lead to C?). If the mentally simulating that in the absence of A, B will not lead to C, an agent may conclude that there is a causal link between A and B.

As counterfactual physical reasoning only gained interest by the AI community recently, there are only a handful of related works, most of which are at the stage of proposing novel counterfactual datasets. CLEVRER [Yi *et al.*, 2020] is a Visual Question Answering (VQA) dataset that was the first to have a counterfactual task for physical reasoning. The counterfactual segment of CLEVRER would ask questions like "without A, which event will not happen?", and the agent would be provided with a video with 2D objects and a few event options. Their proposed Neuro-Symbolic Dynamic Reasoning model that contained a LSTM-based question parser, Mask R-CNN video parser and a graph-based dynamics learning model set the highest benchmark performance for the counterfactual task.

Instead of VQA, Baradel *et al.* [2020] took a different approach of designing counterfactual tasks and proposed the 3D CoPhy dataset. The CoPhy dataset provides an original stream of images, which can reveal information of the confounder dynamics of the scene, after which the initial state is changed via a do-operator and the agent must visually generate the outcome. Their Graph Convolutional Network-based CoPhyNet benchmark was set in their work. Li *et al.* [2020a] tackled the CoPhy benchmark by proposing the graph and deep learning-based Causal World Model that learns unsupervised relationships between the original and alternative outcomes by estimating latent confounding variables. However, they did not provide a direct comparison with CoPhyNEt. Finally, Ates *et al.* [2020] created a purely 2D VQA-based counterfactual dataset that expanded on CLEVRER by providing more complex concepts of 'cause, enable and prevent'.

The second paradigm under causal reasoning, **VoE**, comes from the idea that human infants form expectations of physical events which determines their knowledge of causal links for transformations in physical interactions [Bullock *et al.*, 1982]. Hence, they use these causal links to determine how surprised they are when presented with a plausible or implausible scene. Like counterfactual physical reasoning, **VoE** papers in physical reasoning are comparatively new and all propose their own datasets and approaches. Riochet *et al.* [2018] first proposed the IntPhys dataset that provided 3D (in 2D) scenes of possible and impossible events of 'object permanence', 'shape constancy' and 'continuity'. The goal in **VoE** is to train an agent to recognise the expected video as less 'surprising' than the surprising version. Their convolutional autoencoder and generative adversarial network models performed poorly in comparison with their adult human trials. Piloto *et al.* [2018] also introduced a dataset that showcased additional events of 'solidity' and 'containment', using a variational autoencoder approach to establish a benchmark.

Instead of a purely deep learning approach, Smith *et al.* [2019] used probabilistic simulation along with approximate derendering and particle filtering for the **VoE** task on their own dataset. The model was named ADEPT and performed with high accuracy and even replicated human judgements 'how, when and what' traits of surprising scenes. Finally, Dasgupta *et al.* [2021] proposed a heuristic-based dataset with additional events in support and collision that had augmented metadata of ground-truth features and rules of the physical interaction, representing intermediate stages of reasoning. They showed how a model could potentially leverage on these heuristics to learn with higher accuracy and learn the universal causal relationships in physical reasoning.

4 Challenges and Open Questions

While the earlier sections show that researchers have worked significantly on intuitive physics for machine cognition, we recognise that there still exist multiple challenges in the field.

4.1 Unified Evaluation

We find that there is no agreed upon unified approach to evaluating systems of intuitive physics. Being a broad topic without a 'cookie-cutter' definition, researchers have explored multiple tasks and approaches that one may arguably constitute as 'intuitive physics'. Although researchers have attempted to define their own method of testing intuitive physics, the unification of such methods is lacking, but crucial to creating reliable and verifiable intuitive physics systems. One example is the dataset task that is used for testing models. For instance, many researchers have used the scenario of bouncing balls [Battaglia *et al.*, 2016; Chang *et al.*, 2016; de Avila Belbute-Peres *et al.*, 2018; Xu *et al.*, 2021] for the task of inference, but they created their own version of the challenge and often different metrics for evaluation (Mean Squared Error [Battaglia *et al.*, 2016; de Avila Belbute-Peres *et al.*, 2018] and Accuracy [Chang *et al.*, 2016]). We encourage researchers to work with one version of the task coupled with a standardized metric, so that models by different researchers can be used for direct comparison. If feasible, an explainable metric is preferred like accuracy. One step in this direction may be to create a dataset with a large suite of all challenges widely used by researchers. The closest example is Physion [Bear *et al.*, 2021], which showcases a suite of different tasks on **PIO**. However, such datasets are still needed for other tasks in Figure 2.

4.2 Complex Scenarios

Another challenge in intuitive physics for machine cognition is deploying models in more complex and realistic settings. We define complexity through 3 means. One approach is to focus on intuitive physics using real world datasets. Most datasets (with the exceptions of [Wu *et al.*, 2015; Wu *et al.*, 2016; Wu *et al.*, 2017; Ehrhardt *et al.*, 2019a]) are run on basic and synthetic 2D or 3D physics engine simulations replicating trivial interactions. These synthetic datasets generally do not contain noise that would be found in real-world vision. The second means of added complexity would be to increase the compositionality of interactions in physical reasoning. Datasets from [Dasgupta *et al.*, 2021; Smith *et al.*, 2019; Bear *et al.*, 2021] generally split the events of the interactions into separate videos. For instance, the distinct events of 'barrier' and 'containment' will never be in the same video for the datasets in Dasgupta *et al.* [2021] and Smith *et al.* [2019]. More complex datasets that mix different events of physical interactions pose an additional challenge to reasoning systems as they need to parse the scene's events and reason about them separately. The third means of added complexity is the use of intuitive physical reasoning for embodied tasks [Duan *et al.*, 2021b]. Instead of simply evaluating on videos of static physical interactions, the agents in virtual environments may provide more complex possibilities of deploying intuitive physical reasoning modules for specific tasks. For instance, inference of mass and friction can be useful in the task of pushing an object to a target location.

4.3 Real-world Utility

When considering such potential applications of intuitive physical reasoning systems, one open question we find interesting is **"Specifically, how can intuitive physics systems be deployed in real world applications?"**. Other than allowing machines to learn and think more like humans [Lake *et al.*, 2017], researchers have mentioned the use of intuitive physical reasoning in safe AI systems. There is potential for intuitive physics to be useful for safety applications, but we have not found detailed expositions about achieving this. For instance, how will intuitive physical reasoning systems help autonomous vehicles make more refined decisions by reasoning about interactions in their surroundings? How can physical reasoning modules help robotic manipulators interact with new objects and efficiently learn dynamical parameters? While a plethora of specific applications of intuitive physics exist, these should be discussed more widely and exhaustively. This would not only highlight the importance of intuitive physics research in machine cognition, but it would also encourage researchers to work towards real-world systems that leverage the research.

4.4 Generalizability

As mentioned in Lake *et al.* [2017], one test of a universal physical reasoning system is creating a general-purpose physical simulator that can physically reason in all possible scenarios. This leads us to ask **"How can we create an integrated intuitive physical reasoning system that learns from the context of the scenario?"**. Humans often misjudge their sense of intuitive physics when the context of the interaction is not known [Kubricht *et al.*, 2017]. It would stand to reason that machines too need the context of the physical interaction before making accurate judgments in a complex scenario. A general-purpose intuitive physical reasoning system would need to be able to learn the context of the situation and decide which facet of reasoning it should invoke (*prediction, inference, causal*) and the type of task to tackle. We recognise that this task is extremely challenging, but we hope that researchers recognise the potential of context-driven physical reasoning tasks and systems.

In summary, the challenges and open questions will assist researchers new to the field in converging on several existing and unsolved problems, ranging from dataset development to the definition of new field-specific assessment measures. Table 1 further provides an overview of the field and the categorization of the different facets of physical reasoning, methods, and tasks. Additionally, it provides an overview of current evaluation metrics employed in many parts of physical reasoning. This can assist researchers unfamiliar with the field to select pertinent work for their desired task.

5 Conclusion

To advance towards human-like learning, the integration of intuitive physics and deep learning is crucial [Lake *et al.*, 2017]. We reviewed a range of deep learning papers at the intersection of intuitive physics and AI. These papers were categorised at three levels: facets of intuitive physics, technical approach taken and physical reasoning tasks. While some papers could have multiple labels per categorization, they were organized based on the label most crucial to the deployment of the proposed model. These categorizations give structure to an otherwise amorphous and growing field, while also allowing researchers to swiftly spot gaps in certain areas. Overall, this survey may be used as a starting point in understanding how researchers define intuitive physics for machine cognition, along with challenges and future directions.

References

[Ates *et al.*, 2020] Tayfun Ates, Muhammed Samil Atesoglu, Cagatay Yigit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. Craft: A benchmark for causal reasoning about forces and interactions. *arXiv preprint arXiv:2012.04293*, 2020.

[Baillargeon, 2004] Renée Baillargeon. Infants' physical world. *Current directions in psychological science*, 13(3):89–94, 2004.

[Bakhtin *et al.*, 2019] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *arXiv preprint arXiv:1908.05656*, 2019.

[Baradel *et al.*, 2020] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. In *ICLR*, 2020.

[Bates *et al.*, 2015] Christopher Bates, Peter W Battaglia, Ilker Yildirim, and Joshua B Tenenbaum. Humans predict liquid dynamics using probabilistic simulation. In *CogSci*. Citeseer, 2015.

[Battaglia *et al.*, 2013] 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):18327–18332, 2013.

[Battaglia *et al.*, 2016] Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. *ArXiv*, abs/1612.00222, 2016.

[Bear *et al.*, 2021] Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiau-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

[Bullock *et al.*, 1982] Merry Bullock, Rochel Gelman, and Renée Baillargeon. The development of causal reasoning. *The developmental psychology of time*, pages 209–254, 1982.

[Caramazza *et al.*, 1981] Alfonso Caramazza, Michael McCloskey, and Bert Green. Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition*, 9(2):117–123, 1981.

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

[Chang *et al.*, 2016] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016.

[Dasgupta *et al.*, 2021] Arijit Dasgupta, Jiafei Duan, Marcelo H Ang Jr, Yi Lin, Su-hua Wang, Renée Baillargeon, and Cheston Tan. A benchmark for modeling violation-of-expectation in physical reasoning across event categories. *arXiv preprint arXiv:2111.08826*, 2021.

[de Avila Belbute-Peres *et al.*, 2018] Filipe de Avila Belbute-Peres, Kevin Smith, Kelsey Allen, Josh Tenenbaum, and J Zico Kolter. End-to-end differentiable physics for learning and control. *Advances in neural information processing systems*, 31:7178–7189, 2018.

[Du *et al.*, 2018] Yilun Du, Zhijian Liu, Hector Basevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, and Jiajun Wu. Learning to exploit stability for 3d scene parsing. In *NeurIPS*, pages 1733–1743, 2018.

[Duan *et al.*, 2021a] Jiafei Duan, Samson Yu, Soujanya Poria, Bihan Wen, and Cheston Tan. Pip: Physical interaction prediction via mental imagery with span selection. *arXiv preprint arXiv:2109.04683*, 2021.

[Duan *et al.*, 2021b] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *arXiv preprint arXiv:2103.04918*, 2021.

[Duchaine and Gosselin, 2009] Vincent Duchaine and Clément Gosselin. Safe, stable and intuitive control for physical human-robot interaction. In *2009 IEEE International Conference on Robotics and Automation*, pages 3383–3388. IEEE, 2009.

[Ehrhardt *et al.*, 2019a] Sébastien Ehrhardt, Aron Monszpart, Niloy J Mitra, and Andrea Vedaldi. Unsupervised intuitive physics from past experiences. *arXiv preprint arXiv:1905.10793*, 2019.

[Ehrhardt *et al.*, 2019b] Sébastien Ehrhardt, Aron Monszpart, Niloy Jyoti Mitra, and Andrea Vedaldi. Unsupervised intuitive physics from past experiences. *ArXiv*, abs/1905.10793, 2019.

[Fragkiadaki *et al.*, 2016] Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning visual predictive models of physics for playing billiards. *CoRR*, abs/1511.07404, 2016.

[Gilden and Proffitt, 1994] David L Gilden and Dennis R Proffitt. Heuristic judgment of mass ratio in two-body collisions. *Perception & Psychophysics*, 56(6):708–720, 1994.

[Groth *et al.*, 2018] Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *European Conference on Computer Vision*, pages 724–739. Springer, 2018.

[He *et al.*, 2016] 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.

[Hegarty, 2004] Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, 8(6):280–285, 2004.

[Huang *et al.*, 2018] Siyu Huang, Zhi-Qi Cheng, Xi Li, Xiao Wu, Zhongfei Zhang, and Alexander Hauptmann. Perceiving physical equation by observing visual scenarios. *ArXiv*, abs/1811.12238, 2018.

[Iandola *et al.*, 2016] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[Janner *et al.*, 2018] Michael Janner, Sergey Levine, William T Freeman, Joshua B Tenenbaum, Chelsea Finn, and Jiajun Wu. Reasoning about physical interactions with object-oriented prediction and planning. *arXiv preprint arXiv:1812.10972*, 2018.

[Jaques *et al.*, 2019] Miguel Jaques, Michael Burke, and Timothy Hospedales. Physics-as-inverse-graphics: Unsupervised physical parameter estimation from video. *arXiv preprint arXiv:1905.11169*, 2019.

[Jia *et al.*, 2014] Zhaoyin Jia, Andrew C Gallagher, Ashutosh Saxena, and Tsuhan Chen. 3d reasoning from blocks to stability. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):905–918, 2014.

[Kandukuri *et al.*, 2022] Rama Krishna Kandukuri, Jan Achterhold, Michael Moeller, and Joerg Stueckler. Physical representation learning and parameter identification from video using differentiable physics. *International Journal of Computer Vision*, 130(1):3–16, 2022.

[Kubricht *et al.*, 2017] James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017.

[Lake *et al.*, 2017] 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.

[Lerer *et al.*, 2016] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *International conference on machine learning*, pages 430–438. PMLR, 2016.

[Leslie, 1984] Alan M Leslie. Spatiotemporal continuity and the perception of causality in infants. *Perception*, 13(3):287–305, 1984.

[Li *et al.*, 2016] Wenbin Li, Seyedmajid Azimi, Aleš Leonardis, and Mario Fritz. To fall or not to fall: A visual approach to physical stability prediction. *arXiv preprint arXiv:1604.00066*, 2016.

[Li *et al.*, 2018] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*, 2018.

[Li *et al.*, 2020a] Minne Li, Mengyue Yang, Furui Liu, Xu Chen, Zhitang Chen, and Jun Wang. Causal world models by unsupervised deconfounding of physical dynamics. *arXiv preprint arXiv:2012.14228*, 2020.

[Li *et al.*, 2020b] Yunzhu Li, Toru Lin, Kexin Yi, Daniel Bear, Daniel Yamins, Jiajun Wu, Joshua Tenenbaum, and Antonio Torralba. Visual grounding of learned physical models. In *International conference on machine learning*, pages 5927–5936. PMLR, 2020.

[McCloskey *et al.*, 1983a] Michael McCloskey, Allyson Washburn, and Linda Felch. Intuitive physics: The straight-down belief and its origin. *Journal of experimental psychology. Learning, memory, and cognition*, 9:636–49, 11 1983.

[McCloskey *et al.*, 1983b] Michael McCloskey, Allyson Washburn, and Linda Felch. Intuitive physics: the straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4):636, 1983.

[Mottaghi *et al.*, 2016a] Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. Newtonian scene understanding: Unfolding the dynamics of objects in static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3521–3529, 2016.

[Mottaghi *et al.*, 2016b] Roozbeh Mottaghi, Mohammad Rastegari, Abhinav Kumar Gupta, and Ali Farhadi. ”what happens if...” learning to predict the effect of forces in images. *ArXiv*, abs/1603.05600, 2016.

[Mrowca *et al.*, 2018] Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li Fei-Fei, Joshua B Tenenbaum, and Daniel LK Yamins. Flexible neural representation for physics prediction. *arXiv preprint arXiv:1806.08047*, 2018.

[Piloto *et al.*, 2018] Luis Piloto, Ari Weinstein, Dhruva TB, Arun Ahuja, Mehdi Mirza, Greg Wayne, David Amos, Chia-chun Hung, and Matt Botvinick. Probing physics knowledge using tools from developmental psychology. *arXiv preprint arXiv:1804.01128*, 2018.

[Riochet *et al.*, 2018] 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. *arXiv preprint arXiv:1803.07616*, 2018.

[Runeson *et al.*, 2000] Sverker Runeson, Peter Juslin, and Henrik Olsson. Visual perception of dynamic properties: Cue heuristics versus direct-perceptual competence. *Psychological review*, 107(3):525, 2000.

[Sanborn *et al.*, 2013] Adam N Sanborn, Vikash K Mansinghka, and Thomas L Griffiths. Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, 120(2):411, 2013.

[Smith *et al.*, 2019] Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. 2019.

[Smith *et al.*, in press] KA Smith, AN Sanborn, PW Battaglia, T Gerstenberg, TD Ullman, and JB Tenenbaum. *Probabilistic Models of Physical Reasoning*. in press.

[Song and Boularias, 2018] Changkyu Song and Abdeslam Boularias. Inferring 3d shapes of unknown rigid objects in clutter through inverse physics reasoning. *IEEE Robotics and Automation Letters*, 4(2):201–208, 2018.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[Ullman *et al.*, 2017] 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.

[Watters *et al.*, 2017] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. *Advances in neural information processing systems*, 30:4539–4547, 2017.

[Wu *et al.*, 2015] 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. *Advances in neural information processing systems*, 28:127–135, 2015.

[Wu *et al.*, 2016] Jiajun Wu, Joseph J. Lim, Hongyi Zhang, Joshua B. Tenenbaum, and William T. Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, 2016.

[Wu *et al.*, 2017] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. *Advances in Neural Information Processing Systems*, 30:153–164, 2017.

[Xu *et al.*, 2020] Zhenjia Xu, Zhanpeng He, Jiajun Wu, and Shuran Song. Learning 3d dynamic scene representations for robot manipulation. *arXiv preprint arXiv:2011.01968*, 2020.

[Xu *et al.*, 2021] Kai Xu, Akash Srivastava, Dan Gutfreund, Felix Sosa, Tomer Ullman, Josh Tenenbaum, and Charles Sutton. A bayesian-symbolic approach to reasoning and learning in intuitive physics. *Advances in Neural Information Processing Systems*, 34, 2021.

[Ye *et al.*, 2018] Tian Ye, Xiaolong Wang, James Davidson, and Abhinav Gupta. Interpretable intuitive physics model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018.

[Yi *et al.*, 2020] 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 *ICLR*, 2020.

[Zheng *et al.*, 2013] Bo Zheng, Yibiao Zhao, Joey C Yu, Katsushi Ikeuchi, and Song-Chun Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3134, 2013.

[Zheng *et al.*, 2015] 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.

[Zheng *et al.*, 2018] David Zheng, Vinson Luo, Jiajun Wu, and Joshua B Tenenbaum. Unsupervised learning of latent physical properties using perception-prediction networks. *arXiv preprint arXiv:1807.09244*, 2018.