A Survey on Machine Learning Approaches for Modelling Intuitive Physics

Jiafei Duan∗, Arijit Dasgupta∗, Jason Fischer and Cheston Tan

1Institute for Infocomm Research, A*STAR
2Department of Mechanical Engineering, National University of Singapore
3Department of Psychological and Brain Sciences, Johns Hopkins University
4Centre 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 artificial intelligence, 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
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 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.
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-
<table>
<thead>
<tr>
<th>Facets of Physical Reasoning</th>
<th>Method/Category</th>
<th>Publication</th>
<th>Year</th>
<th>Physical Reasoning Tasks</th>
<th>Evaluation Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>Inverse Rendering</td>
<td>[Mottaghi et al., 2016a]</td>
<td>2016</td>
<td>PI</td>
<td>MPE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Li et al., 2016]</td>
<td>2016</td>
<td>PPI</td>
<td>$F_1$, $R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Leter et al., 2016]</td>
<td>2016</td>
<td>PIO</td>
<td>$\text{Acc}_{\text{MEPE}}$, Log$_L$, Log$_L$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Groth et al., 2018]</td>
<td>2018</td>
<td>PIO</td>
<td>$\text{Acc}_{\text{ME}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Janner et al., 2018]</td>
<td>2018</td>
<td>VSG</td>
<td>$\text{Acc}<em>{\text{MEPE}}$, $\text{Acc}</em>{\text{ME}}$</td>
</tr>
<tr>
<td></td>
<td>Inverse Physics</td>
<td>[Mottaghi et al., 2016b]</td>
<td>2016</td>
<td>PTD</td>
<td>$\text{MEPE}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Ehrhardt et al., 2019b]</td>
<td>2019</td>
<td>PTD</td>
<td>$\text{ME}, \text{Error}$</td>
</tr>
<tr>
<td></td>
<td>Inverse Dynamics</td>
<td>[Watters et al., 2017]</td>
<td>2017</td>
<td>PTD, VSG</td>
<td>$\text{IN}_{\text{L}}$, $\text{MEPE}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Wu et al., 2017]</td>
<td>2017</td>
<td>PPI</td>
<td>$\text{ME}, \text{MAE}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Duan et al., 2021a]</td>
<td>2021</td>
<td>PIO, VSG</td>
<td>$\text{Acc}_{\text{ME}}$, $\text{PSNR}$</td>
</tr>
<tr>
<td></td>
<td>Inverse Physics</td>
<td>[Wu et al., 2015]</td>
<td>2015</td>
<td>PII</td>
<td>$\text{Acc}_{\text{ME}}$, $\text{ME}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Wu et al., 2016]</td>
<td>2016</td>
<td>PII, PTD</td>
<td>$\text{ME}, R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Bataglia et al., 2016]</td>
<td>2016</td>
<td>PII, PTD</td>
<td>$\text{ME}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Chang et al., 2016]</td>
<td>2016</td>
<td>PII</td>
<td>$\text{ME}, \text{Acc}_{\text{ME}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Huang et al., 2018]</td>
<td>2018</td>
<td>Others</td>
<td>$R^2$, $\text{MEPE}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Xu et al., 2021]</td>
<td>2021</td>
<td>PII</td>
<td>$\text{IoU}, \text{ME}$</td>
</tr>
<tr>
<td></td>
<td>Inverse Dynamics</td>
<td>[Fragkiadaki et al., 2016]</td>
<td>2016</td>
<td>PII, PTD</td>
<td>$\text{Error}, \text{Acc}_{\text{ME}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Ye et al., 2018]</td>
<td>2018</td>
<td>PIO, PTD</td>
<td>$\text{Error}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Zheng et al., 2018]</td>
<td>2018</td>
<td>PII, VSG</td>
<td>$R^2$, $\text{MEPE}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Avela Belbute-Peres et al., 2018]</td>
<td>2018</td>
<td>PII, VSG</td>
<td>$\text{ME}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Jaques et al., 2019]</td>
<td>2019</td>
<td>PII, VSG</td>
<td>$\text{SSI}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Kandukuri et al., 2022]</td>
<td>2022</td>
<td>PII</td>
<td>$\text{Error}$</td>
</tr>
<tr>
<td>Causal Reasoning</td>
<td>Inverse Rendering</td>
<td>[Piloto et al., 2018]</td>
<td>2018</td>
<td>VoE</td>
<td>$\text{KD}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Smith et al., 2019]</td>
<td>2019</td>
<td>VoE</td>
<td>$\text{Acc}_{\text{ME}}$, $\text{RMSE}$</td>
</tr>
<tr>
<td></td>
<td>Inverse Physics</td>
<td>[Baradel et al., 2020]</td>
<td>2020</td>
<td>VSG, Others</td>
<td>$\text{ME}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Li et al., 2020a]</td>
<td>2020</td>
<td>VSG, Others</td>
<td>$\text{MRR}, R^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Dasgupta et al., 2021]</td>
<td>2021</td>
<td>VoE</td>
<td>$\text{Acc}_{\text{ME}}$</td>
</tr>
<tr>
<td></td>
<td>Inverse Dynamics</td>
<td>[Riochet et al., 2018]</td>
<td>2018</td>
<td>VoE</td>
<td>$1$-AUC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Aites et al., 2020]</td>
<td>2020</td>
<td>PIO, Others</td>
<td>$\text{Acc}_{\text{ME}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Yi et al., 2020]</td>
<td>2020</td>
<td>PIO, Others</td>
<td>$\text{Acc}_{\text{ME}}$</td>
</tr>
</tbody>
</table>

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: $F_1$ score ($F_1$), modified hausdorff distance (MHD), prediction accuracy ($\text{Acc}_{\text{ME}}$), intersection over union ($\text{IoU}$), log likelihood ($\text{Log}_L$), mean squared error (MSE), prediction variable error ($\text{Error}$), mean euclidean prediction error ($\text{MEPE}$), inverse normalized loss ($\text{IN}_{\text{L}}$), coefficient of determination ($R^2$), structural similarity index (SSI), KL divergence (KD), mean reciprocal rank (MRR) and error rate (1-AUC).

The 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 require 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 inference 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 an 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 of 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 efficiency 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?).

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. Ricochet et al. [2018] first proposed the InfPhys 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 [Beat 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


