Dark, Beyond Deep: A Paradigm Shift to Cognitive AI with Human-like Commonsense

Yixin Zhu[a,*], Tao Gao[a], Lifeng Fan[a], Siyuan Huang[a], Mark Edmonds[a], Hangxin Liu[a], Feng Gao[a], Chi Zhang[a], Siyuan Qi[a], Ying Nian Wu[a], Joshua B. Tenenbaum[b], Song-Chun Zhu[a]

[a] *University of California, Los Angeles*
[b] *Massachusetts Institute of Technology*

Abstract

Recent progress from deep learning is based on a "big data for small task" paradigm, in which massive data is poured into the training of a classifier dedicated to a single task. In this paper, we call for a paradigm shift that flips the data-task relation upside down. Specifically, we propose a "small data for big task" paradigm, wherein a single Artificial Intelligence (AI) system is challenged to develop "commonsense" that can solve a wide range of tasks with small training data. We illustrate the power of this paradigm by reviewing models of commonsense from our groups that synthesize recent breakthroughs from both machine and human vision. We identify functionality, physics, intention, and causality (FPIC), as the four core domains of visual commonsense. FPIC are concerning "why" and "how," which are beyond the dominating "what-and-where" framework of vision. They are invisible in terms of pixels but nevertheless drive the creation, maintenance, and development of visual scenes. Therefore, we coin them as the "dark matter" of vision. Just like our universe cannot be understood by just studying the observable matter, vision cannot be understood without studying FPIC as dark matters. We demonstrate the power of this dark vision approach by showing how to apply FPCI with little training data to solve a wide range of novel tasks including tool-use, planning, utility inference, and social learning in general. In summary, we argue that the next generation of AI must embrace the dark vision approach for solving novel tasks with human-like commonsense.

Keywords: Computer Vision, Artificial Intelligence, Causality, Intuitive Physics, Functionality, Perceived Intention

1. Call for a Paradigm Shift in Vision

Computer vision serves as the front gate to Artificial Intelligence (AI) and a major component of modern intelligent systems. The classic definition of computer vision proposed by the pioneer David Marr [1] is to look "what" is "where." "What" refers to the object recognition (object vision), and "where" denotes the 3D reconstruction and object localization (spatial vision) [2]. Such a definition corresponds to two pathways in the human brain: (1) the Dorsal Pathway for categorical recognition of objects and scenes, and (2) the Ventral Pathway for the reconstruction of depth and shapes, scene layout, visually guided actions, *etc*. This paradigm has guided the geometry-based approaches in the 1980s-1990s and the appearance-based methods in the past 20 years.

Within the past several years, progress has been made in object detection and localization, with the rapid advancement of Deep Neural Networks (DNNs), fueled by massive labelled datasets and hardware accelerations. However, we are still far away from solving computer vision or real machine intelligence; the inference and reasoning abilities of current computer vision systems are narrow and highly specialized, in need of large labeled training data designed for special tasks, and lack of a general *understanding* of how our physical and social world works—common facts that are obvious to an average human adult. To fill in the gap between modern computer vision and human vision, we must look for a broader picture to model and reason about the missing dimensions, which is the human-like commonsense.

By analogy, this is similar to the research in cosmology and astronomy. Physicists proposed a standard cosmology model in the 1980s that the mass-energy observed by electromagnetic spectrum only accounts for less than 5% of the universe, and the rest are dark matters (23%) and dark energy (72%).[1] The properties and characteristics of the dark matters and dark energy have to be reasoned jointly from the visible mass-energy using a sophisticated cosmology model. The dark matters and energy, in return, help to explain the formation, evolution, and motion of the visible universe.

We intend to borrow this physics concept to raise the awareness, in the vision community, of the missing dimensions and the potential benefits of joint representation and joint inference. We argue that humans make such a rich inference from sparse and high-dimensional data and achieve a deep understanding from a single picture because we have common but visually imperceptible knowledge, which can never be recovered with just "what" and "where." Specifically, man-made objects and scenes are designed with latent functionality, determined by the unobservable physical laws and causal relations; see an exam-

*Corresponding author
Email address: yixin.zhu@ucla.edu (Yixin Zhu)

[1] https://map.gsfc.nasa.gov/universe/

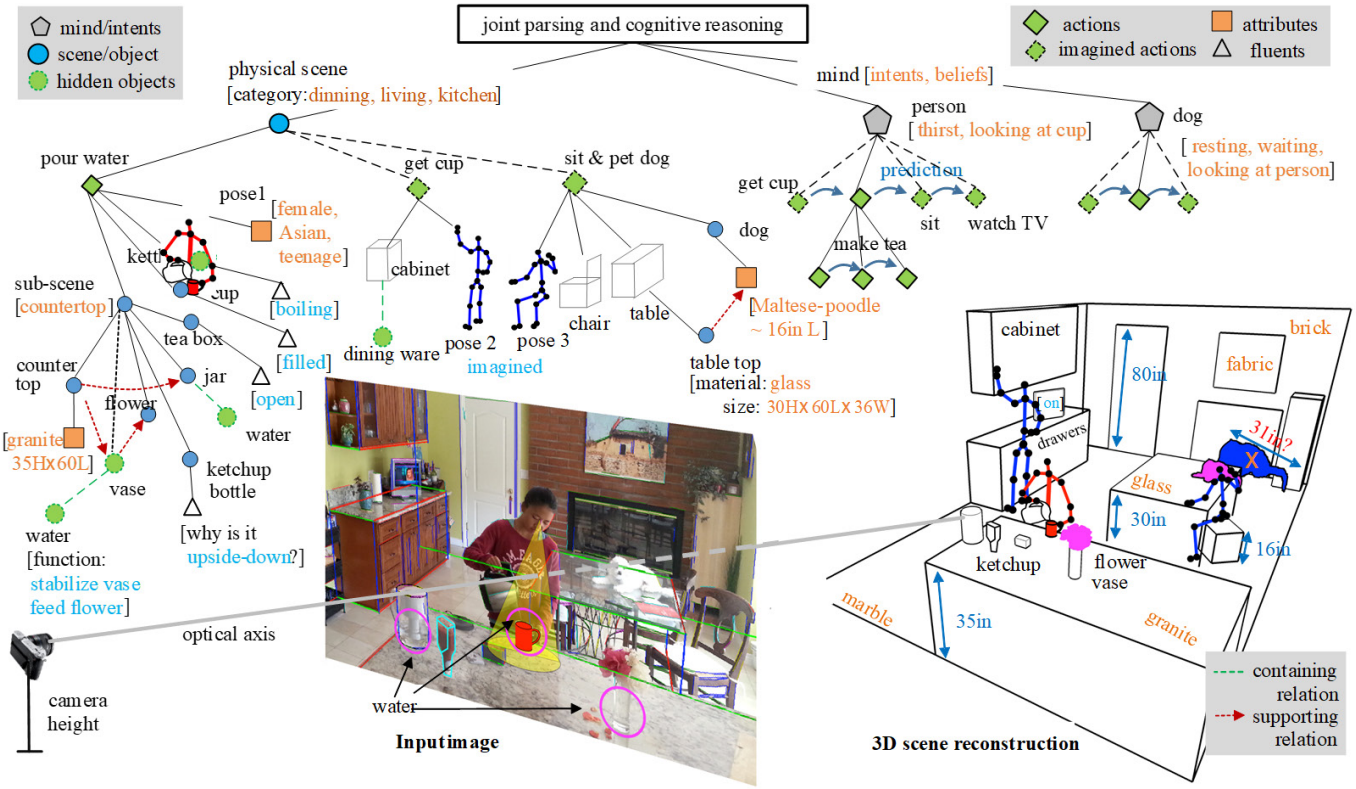


Figure 1: An example of an in-depth understanding of the scene and event by joint parsing and cognitive reasoning. From a single image, a computer vision system should be able to jointly (1) reconstruct the 3D scene, (2) estimate camera parameters, materials, and illumination, (3) parse the scene hierarchically with attributes, fluents, and relations, (4) reason about the intents and beliefs of agents (*e.g.*, the human and dog in this example), (5) predict their actions in time, and (6) recover invisible stuff like water, latent object states, *etc*. We, as humans, can effortlessly (1) predict water is about to come out of the kettle, (2) reason that the intention of putting the ketchup bottle upside down is to utilize gravity for easy use, and (3) there is a hard-to-detect glass table for existing computer vision methods under the dog; otherwise, the parsing results would violate the physical laws as the dog would float in the air. These perceptions can only be achieved by reasoning about the unobservable factors beyond pixels, requiring us to build a commonsense Artificial Intelligence (AI) with human-like core knowledge, which are largely missing in the current computer vision research.

ple in Figure 1. Meanwhile, human activities, especially social activities, are controlled by causality, physics, functionality, social intents, and individual preferences/utilities. In images and videos, many entities (functional objects, fluids, object fluents, intents in mind) and relations (causal effects, physical supports, attraction fields) are impossible to detect by their appearances using existing approaches, and most of these latent factors do not directly appear in pixels. Yet, they are pervasive and governing the placement and motion of visible entities that are relatively easier to detect.

These observations are largely missing in the recent computer vision literature, in which most computer vision tasks have been converted to classification problems, empowered by large-scale annotated data with end-to-end training using neural networks. We call such a paradigm "big data for small tasks."

In this paper, we call for attention to a new promising direction, where "dark entities" and "dark relations" are incorporated into vision and AI research. By reasoning about the unobservable factors beyond visible pixels, we could use only limited data to achieve generalizations to various tasks with human-like commonsense. These tasks are defined as a mixture of "what and where" problems (classification, localization, reconstruction), and "why, how, and what if" problems, including but not limited to physical and social scene understanding, functional reasoning, causal inference, intent prediction, mental state inference, utility learning, tool use, and task planning. We coin this new paradigm "small data for big tasks." Of course, it is well-known that vision is an ill-posed inverse problem [1] where only pixels are seen directly and anything else is hidden/latent. The concept of "darkness" is perpendicular to and richer than the meaning of "latent/hidden" used in vision and probabilistic modeling. It is a measure of the relative difficulty in inferring an entity or a relation from the appearance. One can treat it as a continuous spectrum of "darkness"—from objects like human faces which are relatively easy to detect from appearance and thus considered "visible," to functional objects like chairs which are challenging to recognize from appearance due to its large intraclass variations, and to the entities/relations which are infeasible to recognize by any pixels.

In the remainder of the paper, we start with revisiting a classic view of computer vision in terms of "what" is "where" in Section 2, in which we show that the human vision system is essentially task-driven with its representation and computational mechanism rooted in various tasks. In order to use "small data" to solve "big tasks," we then identify and review four crucial axes of visual commonsense: **F**unctionality, **P**hysics, perceived

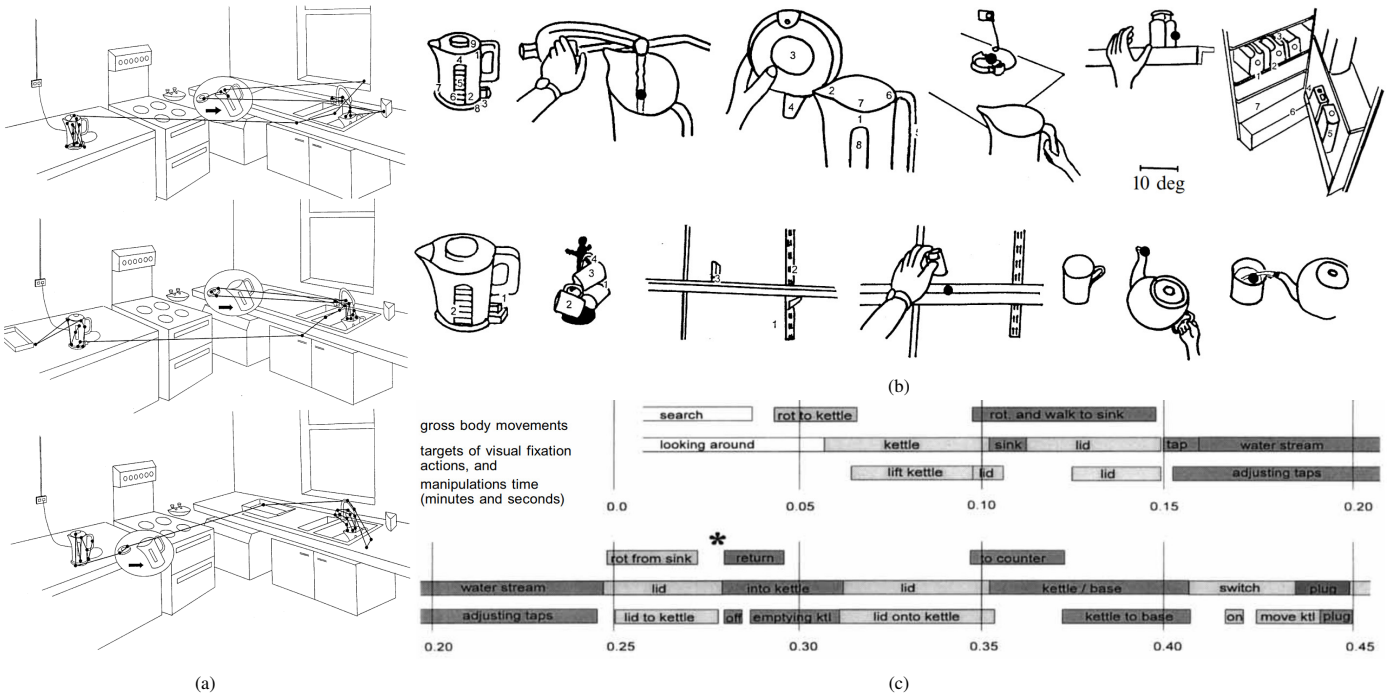


Figure 2: Even for a "simple" task as making a cup of tea, a person can make use of a single vision system to perform a variety of sub-tasks to achieve goals; images adopted from [3]. (a) Record of the fixations made by three different subjects performing the same task of making a cup of tea in a small rectangular kitchen. (b) Examples of fixation patterns drawn from the eye-movement videotape. (c) A sequence of visual and motor events during a single tea-making session.

Intention, and **C**ausality (FPIC). Causality (Section 3) is the basis for intelligent understanding. The application of causality in the physical world (*i.e.*, intuitive physics; Section 4) affords humans the ability to understand the physical world we live in. Functionality (Section 5) is a further understanding of the physical environment when humans intend to interact with the physical world and perform appropriate actions to change the environment to serve human activities. When considering social interactions beyond the physical world, humans need to further infer intention (Section 6) to understand human behavior. In a series of studies, we demonstrate that these four critical aspects of "dark entities" and "dark relations" indeed support various visual tasks beyond just classification tasks. We summarize and discuss our perspectives in Section 7, arguing that it is crucial for the future AI to master these essential ingredients beyond increasing the performance and complexity of data-driven approaches.

2. From Data-driven Vision to Task-driven Vision

What should the vision system afford an agent? From a biological perspective, the majority of the living creatures use a *single* (with multiple components) vision system to perform *thousands* of tasks, in contrast to the current dominating stream in computer vision—a single model designed specifically for a single task. In the literature, such a paradigm to generalize, adapt, and transfer to specific tasks is referred to as the task-centered vision [4]. Given a kitchen as shown in Figure 2, even a simple task like making a cup of coffee consists of multiple sub-goals, including finding objects (object recognition), grasping objects (object manipulation), finding milk in the fridge, and adding sugar (task planning). Prior research has shown that one can finish making a cup of coffee within 1 minute by utilizing a single vision system to facilitate various sub-tasks [3].

Neuroscience studies also suggest similar results, indicating that the human vision system is far more capable than any existing computer vision systems and goes beyond merely memorizing the patterns based on pixels. For example, Fang and He showed that recognizing a face inside an image has a different mechanism compared to seeing an object that can be manipulated as a tool [5]; see Figure 3. Other studies [6] also support the similar conclusion that the images of tool "potentiate" actions even when overt actions are not required in a task. Taking together, these results indicate our biological vision system possesses another mechanism for perceiving object functionality (*i.e.*, how an object can be manipulated as a tool) which is independent of the mechanism in charge of face recognition (and other objects). All these findings call for a quest for the mechanisms of the vision system and natural intelligence.

2.1. 'What': Task-centered Visual Recognition

The human brain can grasp the "gist" of a scene in an image within 200 ms, observed by Potter in the 1970s [8, 9], and Schyns [10] and Thorpe [11] in the 1990s. This line of work often leads researchers to treat categorization as a data-driven process [12, 13, 14, 15, 16], mostly in a feed-forward network architecture [17, 18]. Such thinking has driven the image classification research in computer vision and machine learning

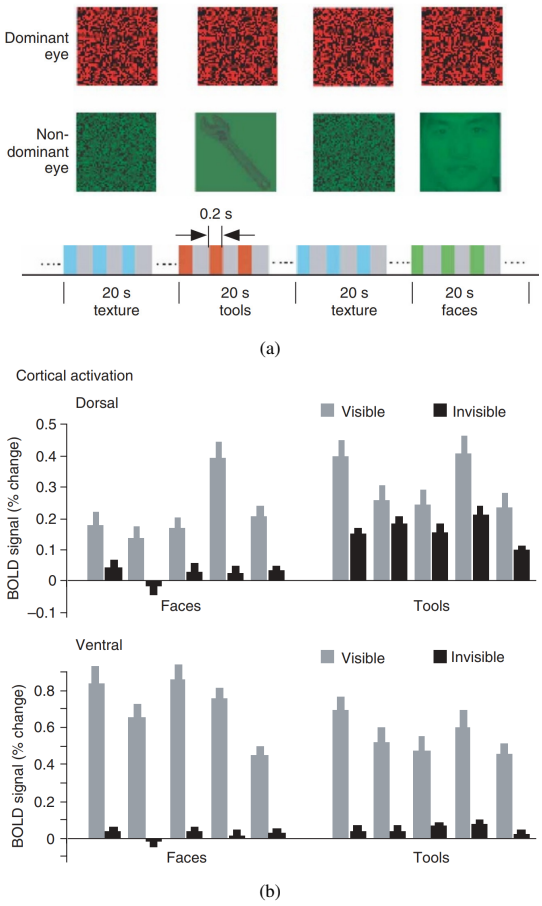


Figure 3: Cortical responses to invisible objects in the human dorsal and ventral pathways [5]. (a) Stimuli (tools and faces) and experimental procedure. (b) Both the dorsal and ventral areas responded to tools and faces. When stimuli were suppressed by high-contrast dynamic textures, the dorsal response remained strong to tools not faces. In contrast, neither tools or faces evoked enough activation in ventral area.

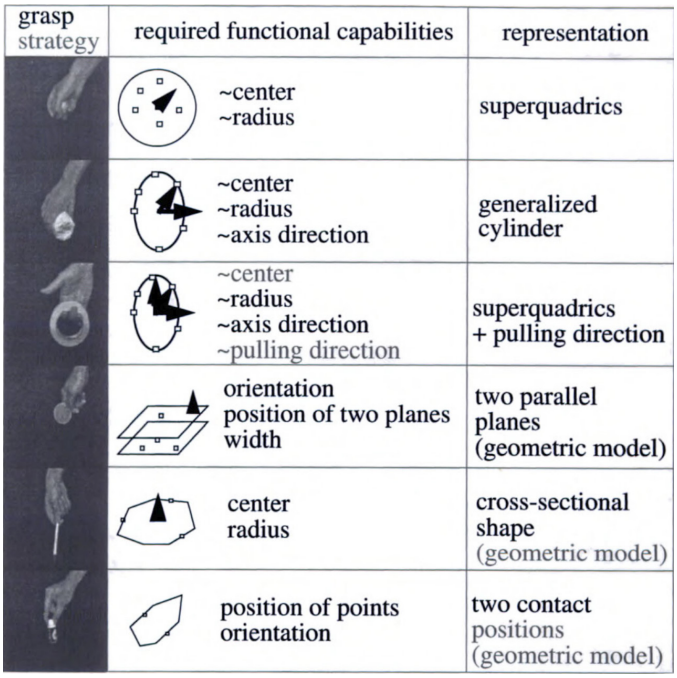


Figure 4: Different grasping strategies require various functional capabilities [7].

in the past decade, and has achieved remarkable progress including the recent success of Convolutional Neural Networks (CNNs) [19, 20, 21].

Despite the fact that these approaches achieved a good performance on scene categorization in terms of the recognition accuracy, a recent large-scale neuroscience study [22] has shown that current DNNs cannot account for image-level behavior patterns of primates (both humans and monkeys), calling for the need for a more precise capture of the neural mechanisms underlying the primate object vision. Furthermore, they have led the focus of scene categorization research away from an important determinant of visual information—the categorization task itself [23, 24]. Simultaneously, these approaches have left it unclear how classification interacts with scene semantics and enables cognitive reasoning. Psychological studies suggest that human vision organizes representations during the inference process even for the "simple" categorical recognition tasks. Depending on a viewer's needs (and tasks), a kitchen can be categorized as an indoor scene, a place to cook, a place to socialize, or specifically as my own kitchen (see Figure 5). As shown in [25], scene categorization and the information gathering process are constrained by these categorization tasks [26, 27], suggesting a bidirectional interplay between the visual input and the viewer's needs/tasks [24]. In addition to the scene categorization, similar phenomena was also found in face recognition [28].

In an early work, Ikeuchi and Hebert [7] proposed a task-centered representation inspired by robotic grasping literature. Specifically, without recovering the detailed 3D models, their analysis suggested that various grasp strategies require the object to afford different functional capabilities, thus the representation of the same object can vary according to the tasks; see Figure 4. For instance, grasping a mug could result in two different grasps—cylindrical grasp of the mug body and the hook grasp of the mug handle. Such findings also suggest that vision (identifying the parts to grasp in this case) is largely driven by tasks; different tasks result in diverse vision representations.

2.2. *'Where': Constructing 3D Scenes in a Series of Tasks*

In literature, computer vision approaches to 3D vision have assumed that the goal is to build an accurate 3D model of the scene through the camera/observer's trajectory. These structure-from-motion and SLAM methods [29] have been the prevailing paradigms in 3D scene reconstruction. In particular, scene reconstruction from a single 2D image is a well-known ill-posed problem; there may exist an infinite number of possible 3D configurations that match the projected 2D observed images [30]. However, the goal here is not to precisely match the 3D ground-truth configuration, but to generate the best possible configuration in terms of functionality, physics, and object relations, in order to enable agents to perform tasks. This line of work has mostly been studied in separation from recogni-

(a) (b) (c)

Figure 5: The experiment presented in [25]. (a) Given an input image of a scene, (b) subjects will categorize the scene as a restaurant if low-pass filters were added to the basic level, and (c) as a cafeteria if added to the subordinate level.

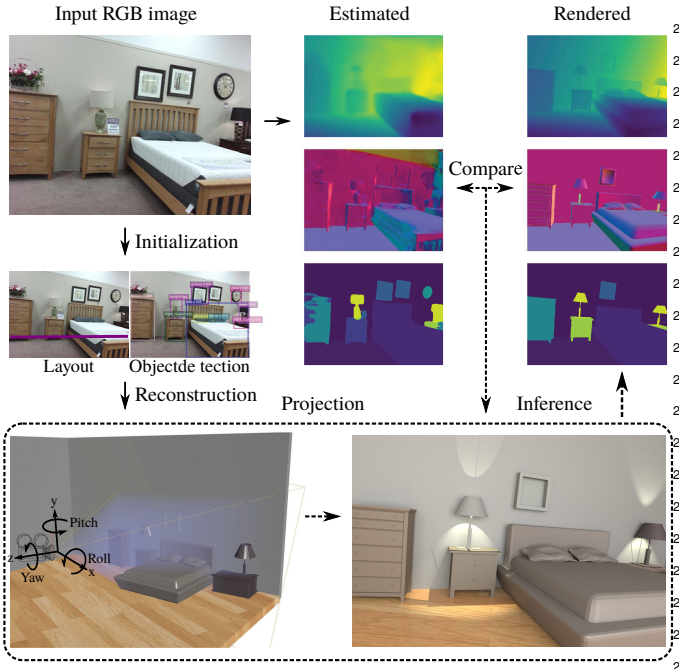


Figure 6: Illustration of the 3D indoor scene parsing and reconstruction in an analysis-by-synthesis fashion [36]. A 3D representation is initialized by individual vision tasks (*e.g.*, object detection, 2D layout estimation). A joint inference algorithm compares the differences between the rendered normal, depth, and segmentation map and the ones estimated directly from the input RGB image, and adjusts the 3D structure iteratively.

tion and semantics until recently [31, 32, 33, 34, 35, 36]; see Figure 6 as an example.

The idea of reconstruction or "cognitive map" has a long history [37]. However, our biological vision system does not rely on such precise computations of features and transformations; there is now abundant evidences that humans represent the 3D layout of a scene in a way that fundamentally differs from any current computer vision algorithms [38, 39]. In fact, multiple experimental studies countenance against global metric representations [40, 41, 42, 43, 44, 45]; human vision is error-proned and distorted in terms of localization [46, 47, 48, 49, 50]. In a case study, Glennerster *et al.* [51] has demonstrated an astonishing lack of sensitivity by observers to dramatic changes in the scale of the environment around a moving observer under various tasks.

Among all the recent evidence, grid cells are perhaps the most well-known discovery to indicate the unnecessity of a precise 3D reconstruction for vision tasks [52, 53, 54]; grid cells encode a cognitive representation of Euclidean space, indicating a different mechanism of perceiving and processing locations and directions. This discovery was later awarded the 2014 Nobel Prize in Physiology or Medicine. Surprisingly, this mechanism not only exists in humans [55], but is also found in mice [56, 57], bats [58], and other animals. Gao *et al.* [59] propose a representational model for grid cells, in which the 2D self-position of the agent is represented by a high-dimensional vector, and the 2D self-motion or displacement of the agent is represented by a matrix that transforms the vector. Such a vector-based model is capable of learning hexagon patterns of grid cells with error correction, path integral, and path planning. A recent study also shows that view-based methods actually perform better than 3D reconstruction-based methods in certain human navigation tasks [60].

Despite these breakthroughs, how we navigate in complex environments while being able to come back to the original place (*i.e.*, homing) remains a mystery in biology and neuroscience. Perhaps, a recent study could shed some light: Vuong *et al.* [61] provide evidence for task-dependent representation of space. Specifically, participants made large, consistent pointing errors that were poorly explained by any single 3D representation. Their study suggests that the mechanism for updating visual directions of unseen targets is neither based on a stable 3D model of the scene nor a distorted one; instead, participants seem to form a flat and task-dependent representation.

2.3. Beyond 'What' and 'Where': Towards Comprehensive Human-like Understanding of the Scenes

Psychological studies have shown that human visual experience is much richer than 'what' and 'where.' As early as in our infancy, humans quickly and efficiently perceive causal relationships (*e.g.*, object launching experiment) [62, 63], agents and intentions (*e.g.*, one entity is chasing another) [64, 65, 66], and the consequences of physical forces (*e.g.*, a precarious stack

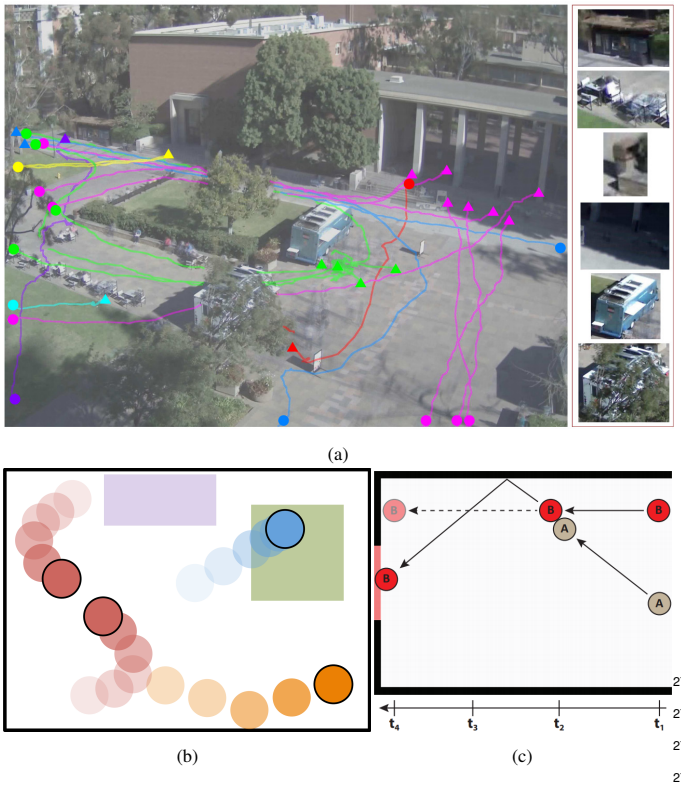


Figure 7: (a) People's trajectories are color-coded by their shared goal destination [69]. The triangles denote destinations, and the dots denote start positions of the trajectories; *e.g.*, people may be heading toward the food-truck to buy food (green), or the vending machine to quench thirst (blue). Due to low resolution, poor lighting, and occlusions, objects at the destinations are very difficult to detect only based on their appearance and shape. (b) An animation reveals the intents, moods and roles of the agents [70]. Motion and interaction of four different pucks moving on a two-dimensional plane are governed by latent physical properties and dynamical laws, such as mass, friction, global and pairwise forces. (c) Intuitive theory and counterfactual reasoning about the dynamics of the scene [71]. Schematic diagram of a collision event between two billiard balls A and B. The solid lines indicated the balls actual movement paths. The dashed line indicates how Ball B would have moved if Ball A had not been present in the scene.

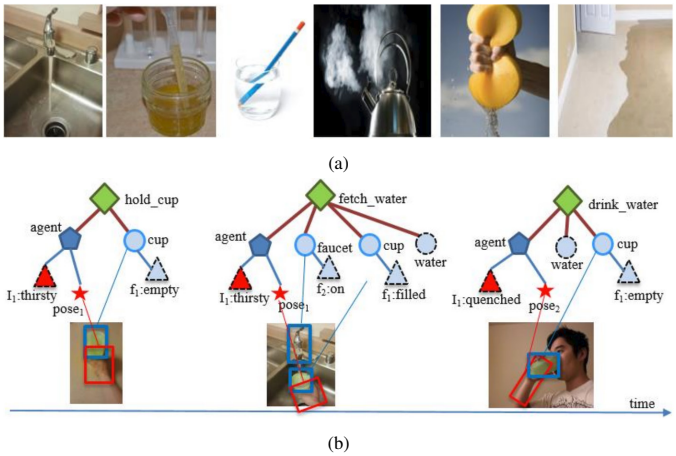


Figure 8: Water, and other fluid, play important roles in our activities, but are hardly detectable in images. (a) Water causes minor appearance changes. (b) The 'dark' entities: water, fluents of the cup and faucet (triangle), and intent of human are shown in dashed nodes. The actions (diamonds) involves agents (pentagon) and cups (object in circles).

of rocks is about to fall in a particular direction) [67, 68]. Rich social and physical concepts can be perceived from both videos [69] and highly impoverished visual inputs [70, 71]; see examples in Figure 7.

To enable an artificial agent with such capabilities, we call for joint reasoning algorithms on a joint representation that integrates (1) the "visible"—traditional recognition categories: objects, scenes, actions, events, *etc*., and (2) the "dark"—higher-level cognition concepts: fluent, causality, physics, functionality, intents, attractions, *etc*. These concepts could be divided into four axes:

I. Fluent and perceived causality. Fluent, a concept coined by Isaac Newton [72, 73] and adopted by AI and commonsense reasoning [74, 75], refers to transient states of objects which are time-variant, such as a cup is 'empty' or 'filled,' a door is 'locked,' a car is 'blinking' to signal a left-turn, and a telephone is 'ringing;' see an example in Figure 8. Such a concept is linked to perceived causality [76] in the psychology literature. Even infants with little exposure to visual experiences have the innate ability to learn causal relationships from daily observation, which leads to a sophisticated understanding of the semantics of the events [77].

Fluents and perceived causality are different from the visual *attributes* [78, 79] of objects. The latter are permanent during the observation, *e.g*., the color of a door, the gender of a person are attributes, not fluents. Some fluents are 'visible,' but many fluents are 'dark.' Human cognition has the innate capability (observed in infants) [77] and a strong inclination to perceive the *causal effects* between *actions* and *change of fluents*; for example, pushing a button causes the light to turn on. Thus, fluents are essential for recognizing actions and understanding the unfolding events. While most vision researches on action recognition have paid a great deal of attention to human poses like walking, jumping, clapping, and to pose-object interactions like drinking and smoking [80, 81, 82, 83], most daily actions, like 'open-door,' are defined by their causes and effects (door fluent changes from 'closed' to 'open,' regardless how it is opened), not by the human poses or spatial-temporal features [84, 85]. Similarly, actions like 'put on clothes,' 'set up a tent' are infeasible to be defined by appearance features due to their complexity, therefore callling for causal reasoning. In fact, the status of a scene can be viewed as a collection of fluents that *record the history of actions*. But as yet, fluents and causal reasoning have not been systematically studied in image understanding, despite their ubiquitous presence in images and videos.

II. Intuitive physics. Psychology studies suggested that approximate Newtonian principles underlie human judgments about dynamics and stability [87, 88]. Hamrick *et al*. [68, 67] showed that the knowledge of Newtonian principles and probabilistic representations is generally applicable for human physical reasoning, and the intuitive physics model is an important perspective for human-level complex scene understand-

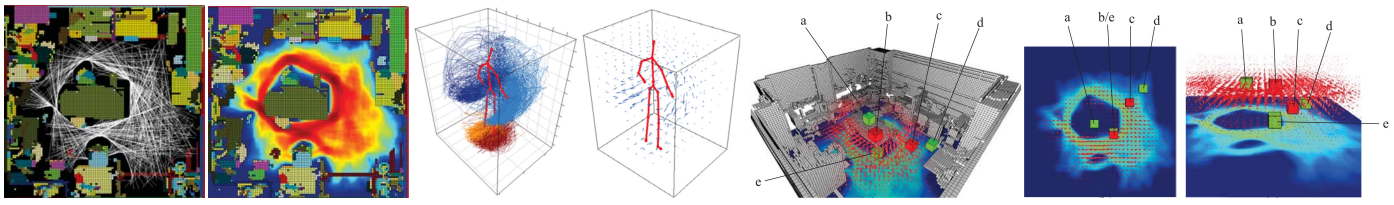


Figure 9: Detecting potential falling objects by inferring human actions and natural disturbance [86]. (a) The hallucinated human trajectories. (b) The distribution of the primary motion space. (c) Secondary motion field. (d) The integrated human action field by convolving primary motions with secondary motions. The objects **a-e** are five typical cases in a disturbance field: the object **b** on edge of a table and the object **c** along the pathway exhibit more disturbances (accidental collisions) than other objects such as **a** in the center of the table, **e** below the table, and **d** on a concave corner in the space.

ing. Other studies have shown that humans are sensitive to objects in a scene that violate certain relations and physical stability [89, 90, 91, 92, 93].

Invisible physical fields govern the layout and placements of objects in a man-made scene. By human design, objects in a scene should be physically stable and safe with respect to gravity and various disturbances [94], such as earthquake, wind/gust, and human activities. Therefore, any 3D scene interpretation or parsing (object and segmentation) must be physically plausible; otherwise they will fall apart [95, 36, 96] (see Figure 9). This observation poses useful constraints for scene understanding and is important for robotics applications [86]. For example, in a rescue or search mission at a disaster relief site, a robot must be able to reason about the stability of an object in the scene and the supporting relations between objects and make cautious moves to maintain stability and safety.

III. Functionality. By fMRI and neurophysiology experiment [97], researchers identified mirror neurons in the premotor cortical area that seem to encode actions through poses and interactions with objects and scenes. Concepts in the human mind are not only represented by prototypes, *i.e.*, exemplars in current vision and machine learning approaches, but also by their functionality [77].

Most man-made scenes are designed to serve multiple human functions, *e.g.*, sitting, eating, socializing, sleeping, *etc.*, and satisfy human needs to an extent comfortable for the functions, *e.g.*, illumination, temperature, ventilation, *etc*. These functions and needs affect the scene layouts [98, 34], the geometric dimensions, the shape of objects, and the selection of materials, but are invisible in images.

IV. Intents and attraction/repulsion fields. Cognitive studies [99] show that humans have a strong inclination to interpret events as a series of goals driven by intents of agents. Such a teleological stance inspired various models in the cognitive literature for intents estimation as an inverse planning problem [100, 101].

We argue that intents can be treated as the transient status of agents (humans and animals), such as being 'thirsty,' 'hungry,' 'tired,' *etc*. They are similar to, but more complex than, the fluents of objects, and come with the following characteristics: (1) They are hierarchically organized in a sequence of goals and are the main factors driving/triggering actions and events in a scene. (2) They are completely 'dark,' *i.e.*, not delegated by any pixels. (3) Unlike the instant change of fluents in response to actions, intents are often formed in long spatio-temporal ranges. For instance, in Figure 7a, when a person is hungry, and sees or knows a food truck in the courtyard, the person decides (intends) to walk to the truck.

During this process, an attraction relation is established in long distance. As it will be illustrated later in this paper, each functional object, such as a food truck, trashcan, or vending machine, emits a field of attraction over the scene, not much different from a gravity field or an electric field. Thus, a scene has many layers of attraction fields or repulsion fields (*e.g.*, odor and grass to avoid) which are completely 'dark,' and a person with a certain intent will move in this field, whose trajectory follows a least-action principle in Lagrange mechanics that derives all motion equations by minimizing the potential and kinematic energies integrated over time.

Reasoning about the intents and attraction fields will be crucial for the following vision and cognition task: (1) Early event and trajectory prediction [102]. (2) Discovering the invisible attractive/repulsive objects and recognizing their functions by analyzing the human trajectories [69]. (3) Understanding scenes by the functions and activities [26]. The attraction fields are longer-range in scene than the functionality map [27, 103] and affordance map [104, 105, 106] studied in the recent literature. (4) Understanding multi-way relations among a group of people and their functional roles [107, 108, 109]. (5) Understanding and inferring the mental states of agents [110, 111].

Summary. Despite their apparent differences at first glance, these four domains do connect with each other in ways that are theoretically important. These connections include: (1) They usually do not easily project onto explicit visual features. (2) Existing computer vision algorithms are neither competent in these domains nor (in most cases) applicable at all. (3) Human vision is nevertheless highly efficient in these domains, and human-level reasoning often builds upon these prior knowledge in these domains.

We argue that the incorporation of these four key elements will advance a vision system in at least three aspects: (1) Generalization. As a higher-level representation, FPIC tends to be globally invariant across the entire human living space. Therefore, knowledge learned in one scene can be transferred to novel situations. (b) Small sample learning. The FPIC encodes essen-

tial prior knowledge for understanding the environment, events, and behavior of agents. As FPIC is more invariant than appearance or geometric features, the leaning of FPIC, which is more consistent and noise-free, is possible even without "big data." (c) Bidirectional inference. Inference with FPIC requires the combination of top-down inference with abstract knowledge and bottom-up inference with visual patterns. The bidirectional process can boost each other as a result.

In the following sections, we discuss these four key elements in greater details.

3. Causal Perception and Reasoning - The Basis for Understanding

Causality is the abstract notion of cause and effect derived from our perceived environment and thus can be used as a prior foundation to construct notions of time and space [112, 113, 114]. People have innate assumptions about causes, and causal reasoning can be activated almost automatically and irresistibly [115, 116]. In our opinion, causality is the pillar for the other three topics (physics, functionality, and intention). For example, an agent must be able to reason about the causes of others' behavior (to understand their intentions) and understand the likely effects of their own actions (to act appropriately). To certain degrees, human understanding depends on the ability to comprehend the causality.

In this section, we start with a brief review of the causal perception and reasoning in psychology, followed by a parallel stream of work in statistical learning. We conclude the section with case studies of causality in computer vision.

3.1. Human Causal Perception and Reasoning

Humans reason about causal relationships through high-level cognitive reasoning. But can we "see" causality directly from vision, just as we see color and depth? In a series of behavioral experiments [117], Scholl's group showed the human visual system can *perceive* causal history by visual commonsense reasoning and represent objects in terms of their underlying inferred causal history—essentially representing shapes by appealing for inferences about 'how they got to be that way.' Inherently, the causal events cannot be directly interpreted merely from vision; they must be interpreted by an agent that understands the distal world [118].

Early psychological work focused on an associative mechanism as the basis for human causal learning and reasoning [119]. During this time, the Rescorla-Wagner model was used to explain how humans (and animals) build expectations using the co-occurrence of perceptual stimuli [120]. However, more recent studies have shown human causal learning is a rational Bayesian process [118, 121, 122] that involves the acquisition of *abstract* causal structure [123, 124] and strength values for cause-effect relationships [125].

The perception of causality is first systematically studied by psychologist Michotte [76] in the context of one billiard ball (A) hitting the other (B). In the classic display, Ball A stops the moment it touches B, and B starts to move immediately with the *same* speed. The visual experience contains not just kinematic motions, but a causal interaction in which A "launches" B. This type of perception has a few notable properties; see [126] for a review:

1. Irresistibility. Even if one is told explicitly that A and B are just patches of pixels that are incapable of mechanical interactions, one is still compelled to perceive launching. One cannot stop seeing salient causality, just as one cannot stop seeing color and depth.
2. Tightly controlled by spatial-temporal patterns of the motions. Just adding a small temporal gap between the stop of A and the motion of B, perceived launching will be destroyed; B's motion will be perceived as self-propelled.
3. Richness. Even the interaction of two balls can support a variety of causal interactions. For example, if Ball B moves with a speed *faster* (vs. the same) than A, then the perception would not be that A "triggers" B's motion. Perceived causality also includes "entraining," which is superficially identical to launching, except that A *continues* to move along with B once they make contact.

Recent cognitive science studies [127] provide more striking evidence showing how deeply causality is rooted in human vision, making the comparison between color and causality more profound. In human vision science, "adaption" is a phenomenon in which an observer adapts to the stimuli after a period of sustained viewing of that stimuli, in a way that perceptual response to the same stimuli becomes weaker. In a particular type of adaption, the stimuli must appear in the same retinotopic position, defined by the reference frame shared by the retina and visual cortex. This type of retinotopic adaption has been taken as a signature of what is a strong evidence of early visual processing of that stimuli. For example, it is well known that the perception of color can induce retinotopic adaption [128]. Strikingly, recent evidence revealed that there is retinotopic adaptation for the perception of causality. After prolonged viewing of a launching display, subsequently viewed displays were judged more often as non-causal, only if the displays are located in the same retinotopic coordinates, which means physical causality is extracted during early visual processing. By using retinotopic adaption as a tool, Kominsky and Scholl [129] recently explored whether launching is a fundamentally different category from *entraining*, in which Ball A moves together with Ball B after contact. The results showed that retinotopically specific adaptation did not transfer between launching and entraining, indicating that there are indeed fundamentally distinct categories of causal perception in vision.

The importance of causal perception is beyond placing labels on different causal events. One unique function of causality is the support of counterfactual reasoning. Observers recruit the capacity of counterfactual reasoning to interpret visual events. In other words, interpretation is not based on what is observed, but on what would have happened but did not happen. In one study [130], participants judged whether one billiard ball caused another one to go through a gate or prevented it from going through. Participants' looking patterns and judgments demonstrated that participants simulated where the target ball would have gone if the candidate cause had been removed from

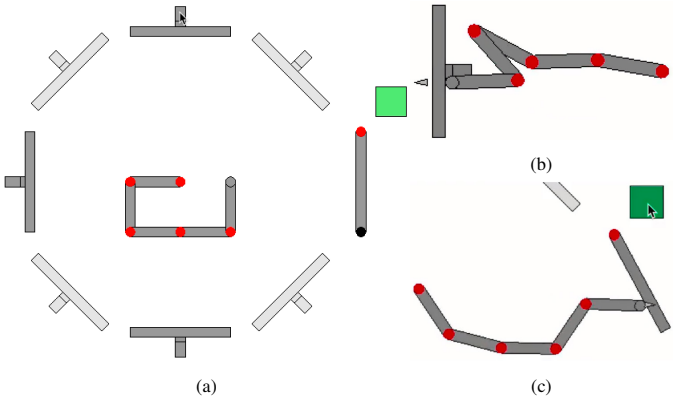
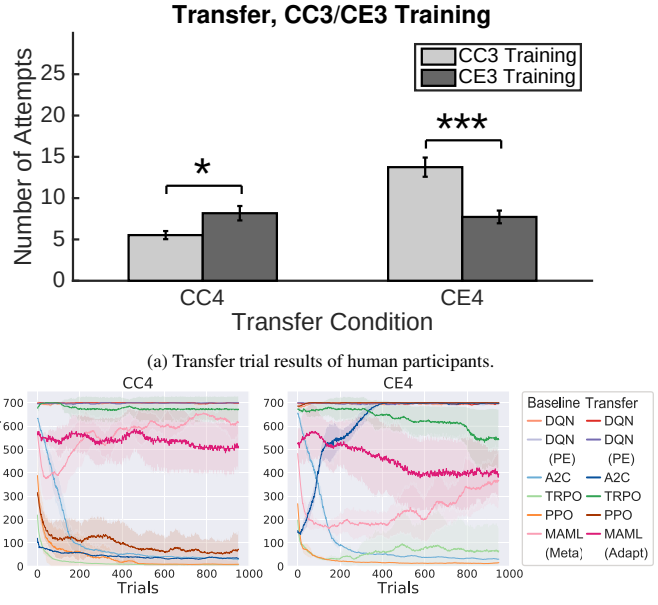


Figure 10: The OpenLock task presented in [124]. (a) Starting configuration of a 3-lever trial. All levers begin pulled towards the robot arm, whose base is anchored to the center of the display. The arm interacts with levers by either *pushing* outward or *pulling* inward. This is achieved by clicking either the outer or inner regions of the levers' radial tracks, respectively. Only push actions are needed to unlock the door in each lock situation. Light gray levers are always locked, which is unknown to both human subjects and Reinforcement Learning (RL) at the beginning of training. Once the door is unlocked, the green button can be clicked to command the arm to push the door open. The black circle located opposite the door's red hinge represents the door lock indicator: present if locked, absent if unlocked. (b) Push to open a lever. (c) Open the door by clicking the green button.



(a) Transfer trial results of human participants.

(b) Transfer trial results of RL agents.

Figure 11: Comparisons between human causal learners and typical RL agents [124]. Common Cause 4 (CC4) and Common Effect 4 (CE4) denote two transfer conditions used in Edmonds *et al*. [124]. (a) Average number of attempts human participants needed to find all unique solutions in the 4-lever common cause (CC4; left) and common effect (CE4; right) conditions. Light and dark grey bars indicate Common Cause 3 (CC3) and Common Effect 3 (CE3) training, respectively. Error bars indicate standard error of the mean. (b) In contrast, RL agents have difficulties transferring learned knowledge to solve similar tasks. Baseline (no transfer) results show the best-performing algorithms (Proximal Policy Optimization (PPO), Trust Region Policy Optimization (TRPO)) achieve approximately 10 and 25 attempts by the end of the baseline training for CC4 and CE4, respectively. Advantage Actor-Critic (A2C) is the only algorithm to show positive transfer; A2C performed better with training for the CC4 condition.

the scene. The more certain participants were that the outcome would have been different, the stronger the causal judgments. These results clearly demonstrated that spontaneous counterfactual simulation played a critical role in scene understanding.

Despite all these evidences demonstrating the important and unique role of causality in human vision, there is in fact much debate in the literature as to whether causal relations are necessary for high-level machine intelligence. Recent successes of systems such as deep Reinforcement Learning (RL) have showcased a broad range of applications [131, 132, 133, 134, 135], the vast majority of which do not learn explicit causal relationships, resulting in a significant challenge for transfer learning in the current dominating machine learning paradigms [136, 137]. One approach to solve such challenging transfer learning problems is to learn a causal encoding of the environment; causal knowledge inherently encodes a transferable representation of the world. Assuming the dynamics of the world are constant, causal relationships will remain true regardless of observational changes to the environment (*e.g*., changing color, shape, position).

Edmonds *et al*. [124] present a complex hierarchical task that requires humans to reason about abstract causal structure. The work proposes a set of virtual "escape rooms" where agents must manipulate a series of levers to open a door and escape from the room; see the illustration in Figure 10. Critically, the task is designed to force agents to form one of the causal structures by requiring agents to find *all* ways to escape from a room, rather than a single way. The work uses 3- and 4-lever rooms and two causal structures: Common Cause (CC) and Common Effect (CE). These causal structures encode different combinations to the room's lock.

After completing a single room, agents are then placed into a room where the perceived environment has been changed, but the underlying abstract causal structure remains the same. In order to reuse the causal structure information acquired in the previous room, the agent needs to learn a mapping between the perception of the new environment and the abstract causal structure on-the-fly. Furthermore, at the end of the experiment, agents are placed in a room with one additional lever; this new room may follow the same or different underlying causal structures to test whether the agent can generalize their acquired knowledge to more complex circumstances.

In this environment, human subjects show a remarkable ability to acquire and transfer knowledge under observationally different but structurally equivalent causal circumstances; see comparisons in Figure 11. Humans approached near-optimal performance and showed positive transfer to rooms with an additional lever. In contrast, recent deep RL methods fail to account for this necessary causal abstraction and show a negative transfer effect. These results suggest current machine learning paradigms do not learn a proper abstract encoding of the environment; *i.e*., they do not learn an abstract causal encoding. Thus, we treat learning causal understanding from perception and interaction as one type of the "dark matters" for current AI systems, one that should be explored further in future work.

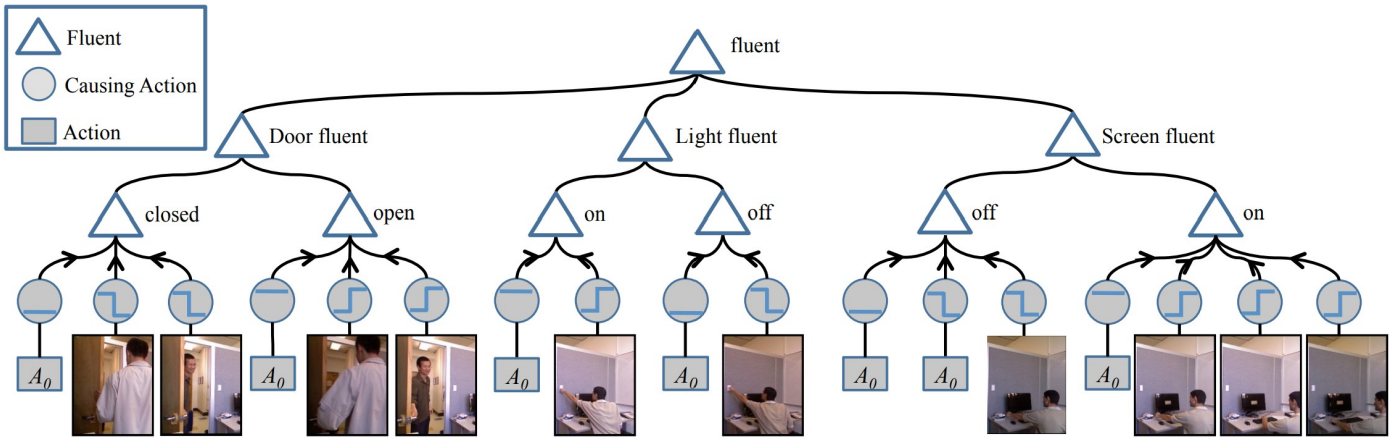


Figure 12: An example of perceptual causality in computer vision [138]. A Causal And-Or Graph for door status, light status, and screen status. Action A_0 represents non-action (a lack of state-changing agent action). Non-action is also used to explain the change of the monitor status to off when the screensaver activates. Arrows point from causes to effects, and undirected lines show deterministic definition.

3.2. Causality in Statistical Learning

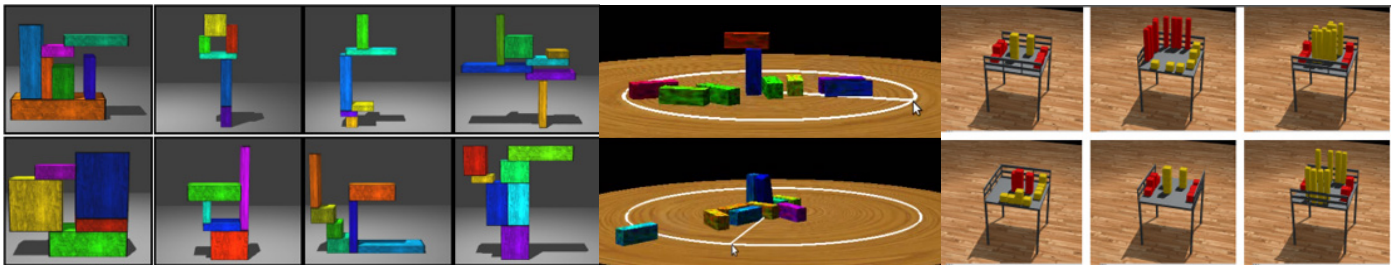
Rubin laid the foundation for causal analysis in his seminal paper [139]; also see [140]. The formulation is commonly called the Rubin causal model. The key concept in Rubin causal model is the potential outcomes. In the simplest scenario where there are two treatments (*e.g.*, smoking or not smoking), for each subject, the causal effect is defined as the difference between the potential outcomes under the two treatments. The difficulty with causal inference is that, for each subject, we only observe the outcome under one treatment that is actually assigned to the subject, and the potential outcome under the other treatment is missing. If the assignment of the treatment to each subject depends on the potential outcomes under the two treatments, a naive analysis by comparing the observed average outcomes of the treatments that are actually assigned to the subjects will result in misleading conclusions. A common scenario for this problem to occur is that there are latent variables that influence both the treatment assignment and the potential outcomes (*e.g.*, a genetic factor that influence both one's tendency to smoke and one's health). A large body of research has been developed to solve this problem. A most prominent example is the propensity score [141], which is the conditional probability of assigning one treatment to the subject given the background variables of the subject. Valid causal inference is possible by comparing subjects with similar propensity scores.

Causality was further developed by Pearl's probabilistic graphical model (*i.e.*, causal Bayesian networks (CBNs)) [142]. CBNs enabled economists and epidemiologists to make inferences for quantities that cannot be intervened in the real world. Typically under this framework, an expert modeler provides the structure of the CBN. The parameters of the model are either provided by the expert or learned from data, given the structure. Inferences are made in the model using the *do* operator, which allows modelers to answer the question *if X is intervened and set to a particular value, how is Y affected*. Concurrently, researchers embarked on a quest to recover casual relationships from observational data [143]. These efforts tried to answer under what circumstances the structure (presence and direction of an edge between two variables in CBN) could be determined from purely observational data [143, 144, 145].

This framework offers a profound tool for fields where real-world interventions are difficult (if not impossible)—such as economics and epidemiology, but lacks many properties necessary for human-like AI. Firstly, despite the attempts to learn causal structure from observational data, most structure learning approaches cannot identify structure beyond a Markov equivalence class of possible structures [145]. Therefore structure learning remains an unsolved problem. Recent work has attempted to tackle this limitation by introducing *active intervention* to enable agents to explore possible directions of undirected causal edges [146, 147]. However, the space of possible structures and parameters is exponential, which has limited the application of CBNs to cases with only a handful of variables. This is partially due to the strict formalism imposed by CBNs, where all possible relations must be considered. Human-like AI should have the ability to constrain the space of possible relations to what is heuristically "reasonable" given the agent's current understanding of the world while acknowledging that such a learning process may not result in the ground truth causal model. That is, we suggest for human-like AI, learners should relax the formalism imposed by CBNs to accommodate significantly more variables without disregarding explicit causal structure (as does the current state of nearly all deep learning models). To make up for this approximation, learners should be in a constant state of active and interventional learning, where their internal causal model of the world is updated with new, confirming, or contradictory evidence.

3.3. Causality in Computer Vision

The classical scientific setting for learning causality in clinical settings consists of Fisher's randomized controlled experiments [148]. Under this paradigm, experimenters control as many confounding factors as possible to tightly restrict their assessment of a causal relationship. While useful for formal science, this is in stark contrast to human ability to perceive causal relationships from observations alone [126, 119, 120].



(a) Will it fall? (b) In which direction? (c) Which is more likely to fall if the table was bumped hard enough, the yellow or the red?

Figure 13: Sample tasks of dynamic scene inferences for physics, stability, and supporting relation presented in [67]. Across a variety of tasks, the Intuitive Physics Engine accounted well for diverse physical judgments in *novel* scenes, even in the presence of varying object properties and unknown external forces that could perturb the environment. This finding supports the hypothesis that human's physics judgment can be viewed as a form of probabilistic inference over the principles of Newtonian mechanics.

These works suggest human causal perception is less rigorous than formal science but still maintains effectiveness in learning and understanding daily events.

Accordingly, computer vision approaches should focus on how humans perceive causal relationships from vision. Fire and Zhu [149, 138] proposed a method to learn causal relationships from image/video input; see an example in Figure 12. Their method pursues causal relations iteratively by asking the same question at each iteration: *given the observed videos and the current causal model, what causal relation should be added to the model to best match the observed statistics of causal events?* To answer this question, the method utilizes the information projection framework [150] by maximizing the amount of information gain after adding a causal relation to the model and then minimizing the divergence between the model and observed statistics.

This method was tested on video datasets consisting of scenes from everyday life: opening doors, refilling water, turning on lights, working at a computer, *etc*. Under the information projection framework, the top-scoring causal relations consistently matched what humans perceived to be a cause in the scene, and low-scoring causal relations matched what humans perceived to not be a cause in the scene. These results indicate the information projection framework is capable of capturing the same judgements made by human causal learners. While computer vision approaches are ultimately observational methods and therefore not guaranteed to uncover the complete and true causal structure, perceptual causality provides a mechanism to achieve human-like learning from observational data.

Causality is crucial for human's video understanding and reasoning, such as tracking humans that are interacting with other objects or the environment, whose visibility might vary over time. Xu *et al*. [151] learn a Causal And-Or Graph (C-AOG) model to tackle such a visibility fluent reasoning problem in tracking interacting objects. They consider the visibility status of an object as a fluent variable, whose change is mostly attributed to the subject's interaction with the surroundings, *e.g*., crossing behind another object, entering a building, or getting into a vehicle, *etc*. The proposed C-AOG can represent the cause-effect relations between an object's visibility fluent and its activities, based on which they develop a probabilistic graphical model to jointly reason the visibility fluent change and track humans in videos. Experimental results demonstrate that with causality reasoning, they can recover complete trajectories of humans in complicated scenarios with frequent human interactions. Xiong *et al*. [152] also defined causality as a fluent change due to a relevant action, and used C-AOG to encapsulate causality learned from human demonstrations in a robot cloth-folding task.

4. Intuitive Physics - Cues of the Physical World

Interacting with the world requires a commonsense understanding of how it operates at a physical level, which does not necessarily require us to precisely or explicitly invoke Newton's laws of mechanics; instead, we rely on intuition, built up through active interactions with the surrounding environment. Humans excel at understanding their physical environment and interacting with objects undergoing dynamic state changes, making approximate predictions from observed events. The knowledge underlying such activities is termed *intuitive physics* [154]. The field of intuitive physics has been explored for several decades in cognitive science and recently reinvigorated by new techniques linked to AI.

Surprisingly, humans develop physical intuitions at an early age [77], well before most other types of high-level reasoning, suggesting the importance of intuitive physics in comprehending and interacting with the physical world. The fact that physical understanding is rooted in visual processing also poses such tasks as important goals for future computational vision systems and AI. In this section, we begin with a short review of intuitive physics in human cognition, followed by recent developments in computer vision and AI by incorporating physics-based simulation and physical constraints for image and scene understanding.

4.1. Intuitive Physics in Human Cognition

Early research in intuitive physics provides several examples of situations where humans demonstrate common misconceptions about how objects in the environment behave. For example, several studies found that humans exhibit striking deviations from Newtonian physical principles when asked to explic-

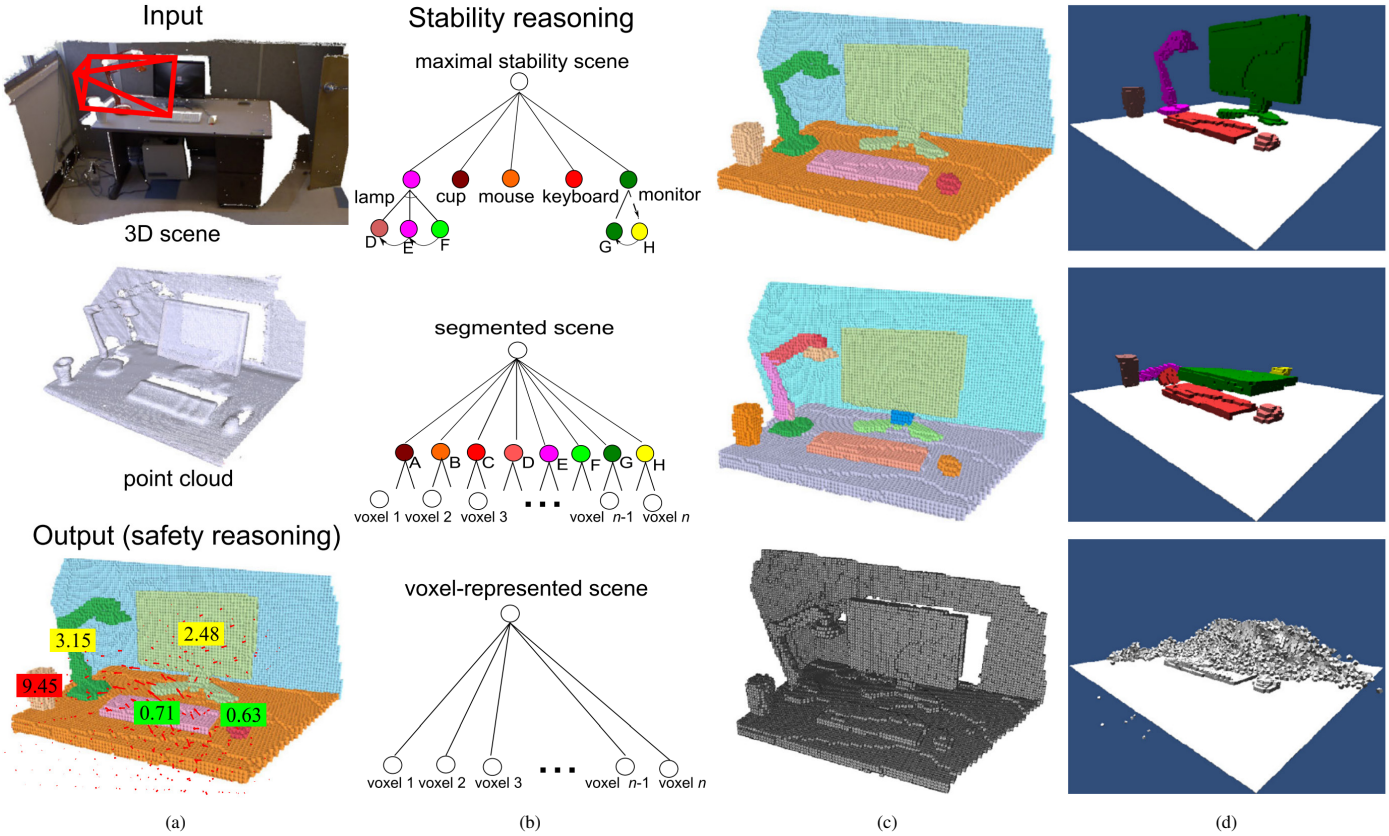


Figure 14: An example explicitly exploiting safety and stability in a 3D scene understanding task [153]. (a) Input: reconstructed 3D scene. Output: parsed and segmented 3D scene as stable objects. The numbers are unsafety scores for each object under the disturbance field (in red arrows) (b) Scene parsing graphs corresponding to 3 bottom-up processes: voxel-based representation (bottom), geometric preprocess including segmentation and volumetric completion (middle), and stability optimization (top). (c) Segmentation result at each step. (d) Physical simulation result of each step.

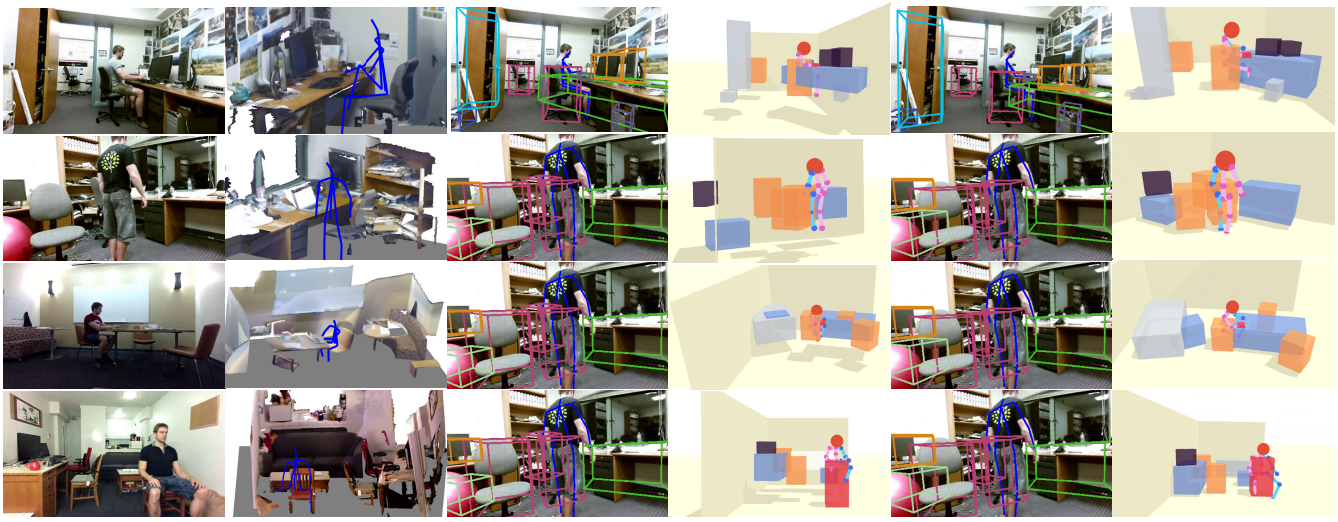
itly reason about the expected continuation of a dynamic event based on a static image representing the situation at a single time point [155, 154, 156]. However, humans' intuitive understanding of physics is much more accurate, rich, and sophisticated than previously expected if *dynamics* and proper *context* were provided [157, 158, 159, 160, 161].

In a glance, humans can perceive whether a stack of dishes will topple, whether a branch will support a child's weight, whether a tool can be lifted, and whether an object can be caught or avoided. In these complex and dynamic events, the ability to perceive, predict, and therefore appropriately interact with objects in the physical world all rely on a rapid physical inference about the environment. Hence, intuitive physics is a core component of human commonsense knowledge and enables a wide range of object and scene understanding.

In an early work [162], Achinstein argued that the brain builds mental models to support inference by mental simulations, analogous to how engineers use simulations for prediction and manipulation of complex physical systems (*e.g.*, analyzing the stability and failure modes of a bridge design before construction). This argument is supported by a recent brain imaging study [163], suggesting that systematic parietal and frontal regions are engaged when humans perform physical inferences even when simply viewing physically rich scenes. Such findings suggest that these brain regions implement a generalized mental engine for intuitive physical inference; *i.e.*, the brain's "physics engine." These brain regions are selective to physical inferences relative to *nonphysical* but otherwise highly similar scenes and tasks. Importantly, these regions are not exclusively engaged in physical inferences, but also overlapped with the parts involved in action planning and tool use, indicating the cognitive and neural mechanisms of understanding intuitive physics have a very intimate relationship with preparing an appropriate action, a critical component linking perception and action

To construct human-like commonsense knowledge, the computational model of the intuitive physics need to be explicitly represented in understanding the agent's environment to support *any* task that involves physics, not particularly adapted to a specific task. This perspective is against the recent "end-to-end" view of AI, in which a neural network directly maps an input image to an output action on a given special task, leaving an implicit internal task representation baked into the weights of the neural network.

Recent breakthroughs in cognitive science provide solid evidence supporting the existence of an intuitive physics module in human scene understanding. Evidence suggests that humans perform physical inferences by running probabilistic simulations in a mental physics engine akin to the 3D physics engines used in video games [165]; see Figure 13. Human intuitive



(a) Input (b) Ground-truth (c) Without physics (d) Without physics (e) With physics (f) With physics

Figure 15: Scene parsing and reconstruction by integrating physics and human-object interactions [164]. Without adding physics, the parsed objects may flow in the air, resulting in unnatural parsing. After adding physics, the parser 3D scene becomes physically stable.

physics can be modeled as an approximated physical engine with Bayesian probabilistic model [67], possessing the following distinguishing properties: (1) physical judgment is achieved by running a coarse and rough forward physical simulation. (2) The simulation is stochastic, which is different from the deterministic and precise physics engine developed in computer graphics. Specific to the tower stability task, there is an uncertainty about the exact physical attributes of the blocks, forming a probabilistic distribution. For every simulation, the model first procedurally samples the blocks' attributes, and then generates predicted states by recursively applying elementary physical rules over short-time intervals as a forward simulation. This process will induce a distribution of simulated results. The stability of a tower is then represented as the probability of tower no-falling in the results. Due to its stochastic nature, this model will judge a tower as stable only when it can tolerate small jitters of its components. This single model fits data from five distinct psychophysical tasks, captures several illusions and biases, and explains core aspects of human mental models and commonsense reasoning that are instrumental to how humans understand their everyday world.

More recent studies have demonstrated that the intuitive physics is not limited to rigid bodies, but also expands to the perception and simulation of liquids [166, 167] and sand [168]. In these studies, the experiments demonstrated that humans do not rely on simple qualitative heuristics to reason about fluid or granular dynamics; instead, they rely on the perceived physical variables to make quantitative judgments. Such results provide converging evidence supporting mental simulation in physical reasoning. For a more in-depth review on intuitive physics in psychology, see [169].

4.2. Physics-based Reasoning in Computer Vision

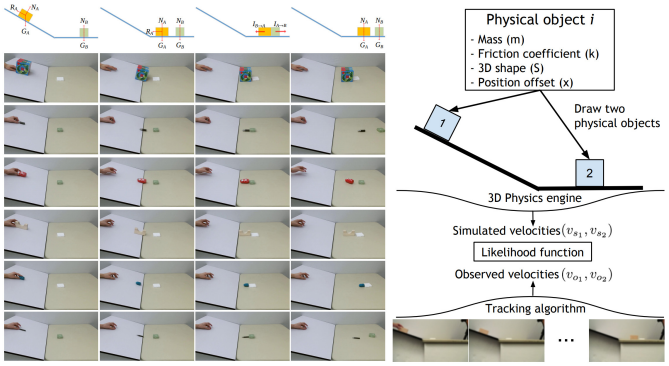
Classic computer vision focuses on appearance and geometric reasoning. Statistical modeling [170] aims to capture the "patterns generated by the world in any modality, with all their naturally occurring complexity and ambiguity, with the goal of reconstructing the processes, objects and events that produced them [171]." Marr conjectured that the perception of a 2D image is an *explicit* multi-phase information process [1], involving (1) an early vision system of perceiving textures [172, 173] and textons [174, 175] to form a primal sketch [176, 177], (2) a mid-level vision system to form 2.1D [178, 179, 180] and 2.5D [181] sketches, and (3) a high-level vision system in charge of full 3D [182, 183, 184]. In particular, he highlighted the importance of different levels of organization and the internal representation [185].

Alternatively, perceptual organization [186, 187] and Gestalt laws [188] aim to resolve the 3D reconstruction problem from a single RGB image without forming the depth cues; but rather using some sorts of priors—groupings and structural cues [189, 190] that are likely to be invariant over wide ranges of viewpoints [191], resulting in feature-based approaches [192, 84].

However, both approaches have well-known difficulties resolving the appearance [193] and geometric ambiguities [29]. To address this challenge, incorporating physics with modern computer vision systems and methods have demonstrated a few interesting and successful cases that dramatically improved the performance of the traditional appearance or geometry-based methods. In certain cases, the ambiguity has been shown to be extremely difficult to resolve by the current state-of-the-art data-driven classification methods, indicating the significance of the physical cues during the perception of our daily environments; see examples in Figure 15.

Through modeling and adapting the physics into computer vision algorithms, the following two problems have been broadly studied:

1. Stability and safety in scene understanding [94]. This line of work is mainly based on a simple but crucial observa-



(a) Snapshots of datasets (b) Galileo model

Figure 16: Inferring the dynamics of the scenes [194]. (a) Snapshots of the dataset. (b) Overview of the Galileo model that estimates physical properties of objects from visual inputs by incorporating the feedback of a physics engine in the loop.

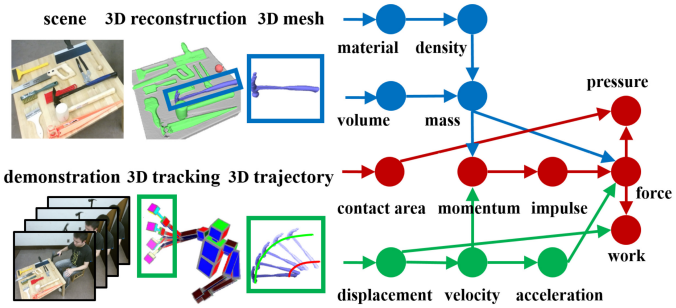


Figure 17: Thirteen physical concepts involved in tool-use and their compositional relations [209]. By parsing human demonstration, the physical concepts of material, volume, concept area, and displacement are estimated from 3D meshes of tool (blue), trajectories of tool-use (green) or jointly (red). The higher-level physical concepts can be further derived recursively.

tion in man-made environments: by human design, objects in static scenes should be stable in the gravity field and be safe with respect to various physical disturbances. Such an assumption poses key constraints for a physically plausible interpretation in scene understanding.

2. Physical relations in the 3D scenes [36]. Humans excel in reasoning the physical relations in the 3D scene such as supporting, attaching, and hanging. Those relations represent a deeper understanding of the 3D scenes beyond the observed pixels that could benefit various applications in robotics, virtual reality, and augmented reality.

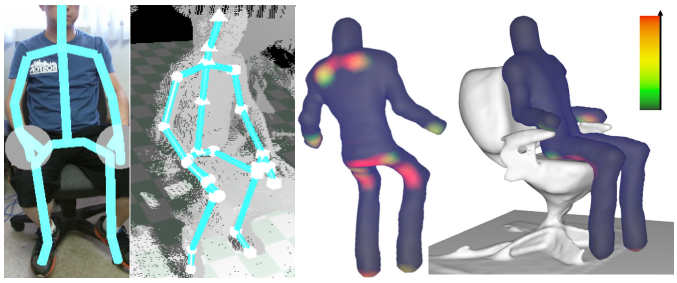
The ideas to incorporate physics to address vision problems could be traced back to Helmholtz and his argument for "unconscious inference" as part of the formation of visual impressions, whose function is to infer the probable causes of sensory input [195]. The very first formal solution in computer vision dates back to Roborts' solutions to completely parse and reconstruct 3D block world in 1963 [196], which inspired later researchers to realize the importance of the violation of physical laws for scene understanding [197] and the stability in generic robot manipulation tasks [198, 199].

Integrating physics for scene parsing and reconstructions was revisited in 2010s with modern computer vision systems and methods. From a single RGB image, Gupta *et al*. proposed a qualitative physical representation for indoor [31, 98] and outdoor [200] scenes, where the algorithm infers the volumetric shapes of objects and relationships (occlusions and supporting relations) describing 3D structure and mechanical configurations. In the next few years, other work [201, 202, 203, 204, 205, 105, 32, 206, 207, 34] also integrated a component for inferring the physical relations such as supporting relations for various scene understanding tasks. More recently, Liu *et al*. [35] inferred physical relations for joint semantic segmentation and 3D scene reconstructions for outdoor scenes. Huang *et al*. modeled the support relations as edges in the human-centric scene graphical model, inferred the relations by minimizing the supporting energies among objects and room layout [36], and enforced the physical stability and plausibility by penalizing the intersection between reconstructed 3D objects and room layout [96, 164].

The aforementioned recent work mostly adopts simple physics cues; *i.e*., no or very limited physics-based simulation is applied. The first recent work that utilizes an actual physics simulator with modern computer vision methods was proposed by Zheng *et al*. in 2013 [153, 86, 94]. As shown in Figure 14, the proposed method first groups the primitives to physically stable objects by optimizing the stability and scene prior, and then predicts the unsafety scores by inferring the hidden and situated causes (disturbance fields), resulting in a physically plausible scene interpretation (voxel segmentation). This line of work is further explored by Du *et al*. [208] by integrating an end-to-end trainable network and synthetic data.

Going beyond stability and supporting relations, Wu *et al*. [194] integrated physics engines with deep learning to predict future dynamical evolution of static scenes. Specifically, a generative model named Galileo is proposed for physical scene understanding based on real-world videos and images. As shown in Figure 16, the core of the generative model is a 3D physics engine, operating on an object-based representation of physical properties, including mass, position, 3D shape, and friction. The model can infer these latent properties using relatively brief runs of Markov Chain Monte Carlo (MCMC), which drive simulations in the physics engine to fit key features of visual observations. They further explore directly mapping visual inputs to physical properties, inverting a part of the generative process using deep learning [210]. Object-centered physical properties like mass, density, and coefficient of restitution from unlabeled videos could be directly derived across various scenarios. With a new dataset named *Physics 101* containing 17,408 video clips and 101 objects of various materials and appearances (shapes, colors and sizes), the proposed unsupervised representation learning model, which explicitly encodes basic physical laws into the structure, can learn physical properties of objects from videos.

Integrating physics and predicting the future dynamics open up quite a few interesting directions in computer vision. For example, given a human motion or demonstration of executing a task as a RGB-D image sequence, Zhu *et al*. [209] cal-



(a) Input (b) Inferred forces on body parts

Figure 18: Inferring forces from videos [211]. (a) The stick-man model captured using a Kinect sensor. It is first converted into a tetrahedralized human model and then segmented into 14 body parts. (b) Using FEM simulation, the forces are estimated at each vertex of the FEM mesh.

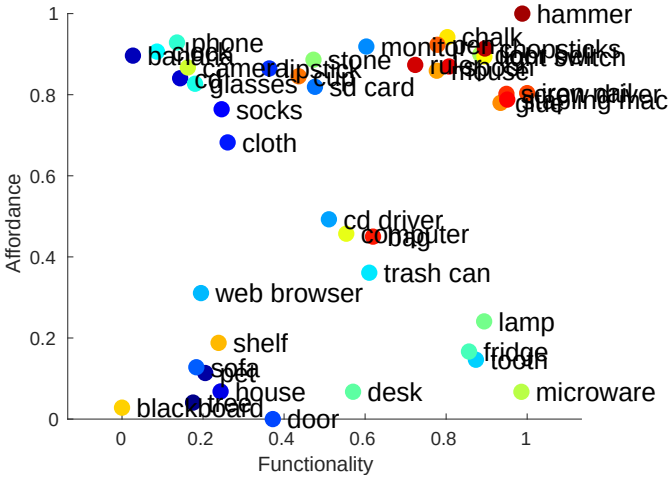


Figure 19: The likelihood of an daily object to be used as a tool with respect to its functionality and affordance. The hotter the color is, the higher the probability is. The functionality score is the average response of "Can it be used to change the status of another object?", and the affordance score is the average response of "Can it be manipulated by hand?"

culated various physical concepts merely from a single example of tool-use (see Figure 17), enabling its ability to reason about the essential physical concepts in the task (*e.g.*, forces in cracking nuts). As the fidelity and the complexity of the simulation increase, Zhu *et al.* [211] were able to infer the forces during human sitting behavior, resulting in estimated force on various body mesh using a Finite Element Method (FEM); see Figure 18.

Physics-based reasoning not only can be applied for the above scene understanding tasks, but also have been successfully demonstrated in human pose and hand recognition and analysis tasks. For example, Brubaker *et al.* [212, 213, 214] estimated contact forces and internal joint torques of human actions using a mass-spring system. Pham *et al.* [215] attempted to infer hand manipulation forces during human hand-object interactions. In computer graphics, soft body simulation has been used to jointly track human hands and calculate contact forces from videos [216, 217].

5. Functionality and Affordance - The Possibility for Task and Action

The perception of the environment inevitably leads to some course of action [218, 219]; Gibson argued that the clues to indicate the opportunity for actions in the nearby environment are perceived in a *direct*, *immediate* way with no sensory processing. It is particularly true for man-made objects and environment, as "an object is first identified as having important functional relations" and "perceptual analysis is derived of the functional concept" [220]; for instance, switches for flipping, buttons for pushing, knobs for turning, hooks for hanging, caps for rotating, handles for pulling, levers for sliding, *etc.* These arguments are the central piece of the Affordance Theory [221] which is based on Gestalt theories and has a significant influence in changing the way we consider visual perception and scene understanding.

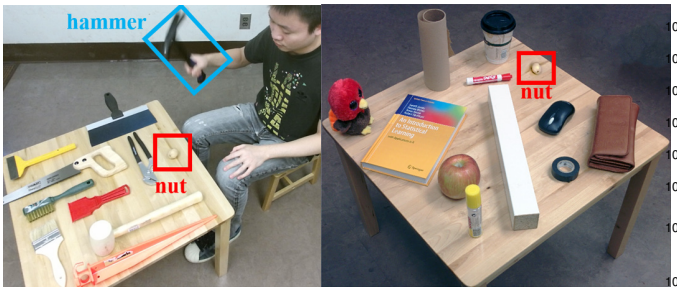
Similarly, functional understanding of objects and scenes relates to identifying the possible set of tasks which can be performed with an object [222]. In contrast to affordances which are directly dependent on the actor, functionality is a permanent property of an object independent of the characteristics of the user. These two interweaving concepts are more invariant concepts for object and scene understanding than their geometry and appearance dimensions. Specifically,

1. Objects, especially man-made ones, are defined by their functions and actions that they are involved with.
2. Scenes, especially man-made ones, are defined by the activities that can be performed in the scenes.

Functionality and affordance are interdisciplinary topics and have been reviewed from different perspectives in the literature (*e.g.*, [223]). In this section, we start with a case study of tool-use in animal cognition to motivate the importance of incorporating functionality and affordance for AI and computer vision, which has been mostly ignored in the literature. A review of functionality and affordance in computer vision is provided, from both the object-level and scene-level. In the end, we review some recent manipulation literature in robotics that focuses on identifying the functionality and affordance of the objects, which is complement to the previous reviews in data-driven approaches [224] and affordance tasks [225].

5.1. Revelation from Tool-use in Animal Cognition

The ability to use an object as a tool to alter another object to accomplish a task has traditionally been regarded as an indicator of the intelligence and complex cognition [226, 227]. Researchers have been using tool-use as the hallmark of the human intelligence to separate humans from non-human animals [228], until relatively recently that Dr. Goodall observed wild chimpanzees manufacture and use tools with regularity [229, 230, 231]. In addition to chimpanzees, studies have been reported on tool-uses by other species. For example, Santos *et al.* [232] trained two species of monkeys on a task to choose one of two canes to reach food under various conditions that involve different types of physical concepts (*e.g.*, materials, connectivity, gravity). Hunt *et al.* [233] and Weir *et al.* [234] reported that New Caledonian crows can bend a piece of straight



(a) Learning (b) Inference

Figure 20: Finding proper tool candidate in novel situations [209]. (a) In a learning phase, a rational human is observed picking a hammer among other tools to crack a nut. (b) In an inference phase, the algorithm is asked to pick the best object (*i.e.*, the wooden leg) on the table for the same task. This generalization entails the reasoning about functionality, physics, and causal relations among the objects, actions, and tasks.

wire into a hook and successfully use it to lift a bucket containing food from a vertical pipe.

These discoveries suggest that some animals can reason about the *functional* properties, physical concepts, and causal relations of tools given a specific task using domain general mechanisms, despite the large differences of their visual appearance and geometry features. Tool-use is of particular interests and poses two major challenges in comparative cognition [235], which also hinders the reasoning ability in computer vision and robotics systems.

First, why can some species come up with innovative solutions while others cannot when facing the same situations? See an example in Figure 20. By observing only a single demonstration of a person achieving a complex task—cracking a nut; we humans can effortlessly reason about the potential candidates capable of completing the same task from another set of random and completely different objects, despite the large visual differences. Such a large intra-class variance demonstrated in reasoning about the tool-use is extremely difficult to capture and resolve in the modern computer vision systems. Without a consistent visual pattern, it is a long-tail problem for a visual recognition challenge; in fact, the very same object can serve for multiple functions depending on context. The type or category of the object is no longer bound by its conventional object name (*i.e.*, a hammer); instead, it is defined by its functionality (*e.g.*, it has the *function* to crack a nut or open a bottle of beer).

Second, what does it take for such a capability to emerge if one does not posses such a reasoning capability? For example, New Caledonian crows are well-known for their propensity and dexterity at making and using tools. But a distantly related cousin, the rooks, are able to reason and use the tools in the lab setting, even they do not use tools in the wild [236]. These findings suggest that the ability to represent tools may be a more domain-general cognitive capacity on reasoning about functionality rather than an adaptive specialization.

5.2. Perceiving Functionality and Affordance

> "The theory of affordances rescues us from the philosophical muddle of assuming fixed classes of objects, each defined by its common feature and then give a name .... You do not have to classify and label things in order to perceive what they afford .... It is never necessary to distinguish all the features of an object and, in fact, it would be impossible to do so."
>
> — J. J. Gibson, 1977 [221]

The idea to incorporate functionality and affordance into computer vision and AI could be dated back to the second IJCAI conference in 1971 by Freeman and Newell [237], in which they argued that available structures should be described in terms of functions provided and functions performed. The concept of affordance is later coined by Gibson [221]. Based on the classic geometry-based "arch-learning" program [238], Winston *et al*. discussed the use of function-based descriptions of object categories [239]. They pointed out that one can use a single functional description to represent all possible cups, despite there could be an infinity of individual physical descriptions for objects like "cups." In their "Mechanic's Mate" system [240], Brady *et al*. proposed a semantic net descriptions based on 2-D shapes together with a generalized structural description [241]. "Chair" and "Tool," exemplar categories researchers used for studies in functionality and affordance, were first systematically discussed with a computational method by Ho [242] and DiManzo *et al*. [243], respectively. Inspired by the functional aspect of the "chair" category in Minsky's book [244], the first work that uses purely functional-based definition of an object category (*i.e.*, no explicit geometric or structural model) was proposed by Stark *et al*. [245]. These ideas of integrating functionality and affordance with computer vision systems have been modernized in the past decade; below, we review some representative work.

"Tool" is of particular interests in computer vision and robotics, partly due to its nature to change *other* objects' status. Motivated by the studies of tool-use in animal cognition, Zhu *et al*. [209] cast the tool understanding problem as a *task-oriented* object recognition problem, which aims at understanding the underlying functions, physics, and causality in using objects as "tools." As shown in Figure 21, a tool is a physical object used in human action to achieve the task, such as a hammer or a brush. From this new perspective, any objects can be viewed as a hammer or a shovel, and this generative representation allows computer vision algorithms to generalize object recognition to novel functions and situations by reasoning about the underlying mechanisms in various tasks, and go beyond memorizing typical examples for each object category as the prevailing appearance-based recognition methods do in the literature. Combined both the physical and the geometry aspects, Liu *et al*. [246] further learned the physical primitive decomposition for tool recognition and tower stability test. Using a more data-driven fashion, Fang *et al*. [247] extracted object affordance from labeled demonstration videos.

"Container" is ubiquitous in daily life and considered as a half-tool [248]. The study of containers can be traced back to a series of studies by Inhelder and Piaget in 1958 [249], in which they showed six-year old children could still be confused by the

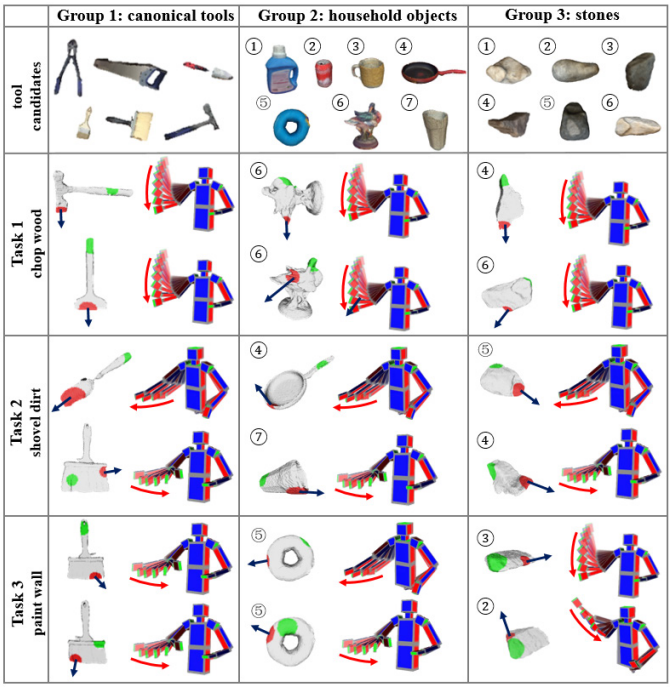


Figure 21: Given three tasks: chop wood, shovel dirt, and paint wall, the algorithm proposed by Zhu *et al*. [209] picks and ranks objects for each task among objects in three groups: (1) conventional tools, (2) household objects, and (3) stones, and output the imagined tool-use: affordance basis (the green spot to grasp with hand), functional basis (the red area applied to the target object), and the imagined action pose sequence.

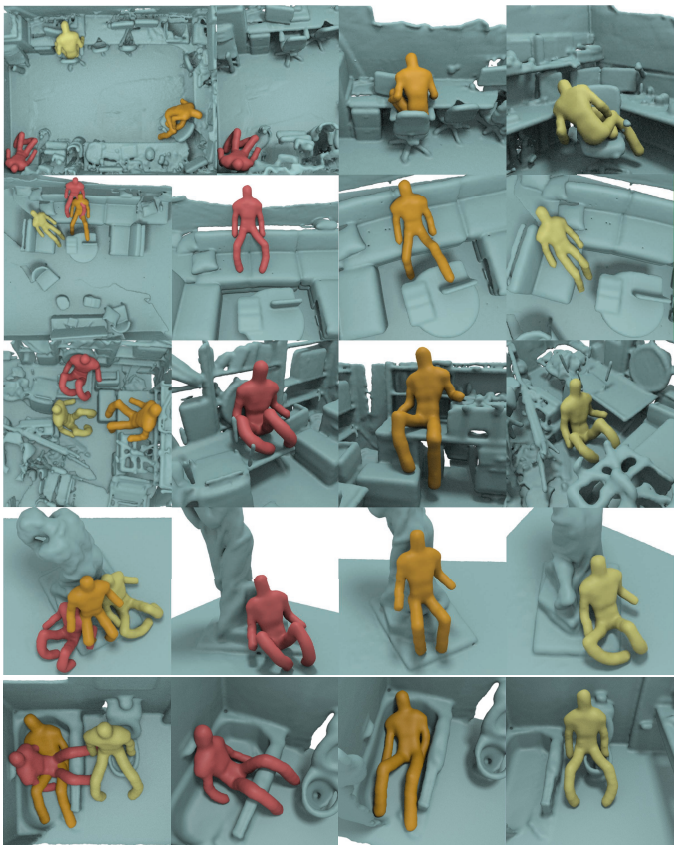


Figure 22: (a) Top 3 poses in various scenes for affordance (sitting) recognition [211]. Zoom-in views of the (b) best, (c) second, (d) third choice of the sitting poses. Top two rows are canonical scenarios, middle row is cluttered scenario, and the bottom two rows are novel scenarios, which demonstrated a large generalization and transfer capability.

complex phenomenon of pouring liquid into containers. Container and containment relations are of particular interests in AI, computer vision, and psychology due to the fact that it is one of the earliest spatial relations to be learned, preceding other common relations (*e.g*., occulusions [250] and support relations [251]). As early as 2.5 months old, infants can already understand containers and containment relations [252, 253, 254]. In AI community, researchers have been adopting common-sense reasoning [255, 256, 257] and qualitative representation/reasoning [258, 259] for reasoning about container and containment relation, mostly focusing on ontology, topology, first-order logic, and knowledge base.

More recently, physical cues have demonstrated a strong capability of facilitating the reasoning about functionality and affordance in container and containment relation. For instance, Liang *et al*. [260] demonstrated that using physics-based simulation is more robust and transferable in identifying containers compared with using features extracted by appearance and geometry cues in three tasks—"What is a container?", "Will an object contain another?", and "How many objects will a container hold?" This line of research accords with the recent findings of intuitive physics in psychology [67, 158, 166, 167, 168, 169], which also enabled a few interesting directions and applications in computer vision, including reasoning about liquid transfer [261, 262], container and containment relation [263], and object tracking [264].

"Chair" is an exemplar class for affordance; the latest studies on object affordance include reasoning about both geometry and function, thereby achieving better generalizations to unseen instances than conventional appearance-based machine learning approaches. In particular, Grabner *et al*. [104] designed an "affordance detector" for chairs by fitting typical human sitting poses to 3D objects. Going beyond visible geometric compatibility, through physics-based simulation, Zhu *et al*. [211] inferred the forces/pressures on various body parts while sitting on a chair; see Figure 22. Thus, their system is able to "feel," in numerical terms, discomfort when the forces/pressures on body parts exceed comfort intervals.

"Human" context is proven to be a critical component in modeling constraints among objects in a scene, in addition to recognizing chairs. A fundamental reason is that man-made scenes are functional spaces that serve human activities, and most objects in the indoor scenes are functional entities that assist human actions [221]. At the object-level, by learning human-object relations, Jiang *et al*. proposed methods to learn object arrangment [265] and object labelling [106] using human context. At the scene-level, Zhao *et al*. [34] modeled the functionality of the 3D scenes as the compositional and contextual relations within the scene. To further explore the hidden human context in the 3D scenes, Huang *et al*. [36] propose a stochastic method to parse and reconstruct the 3D scene

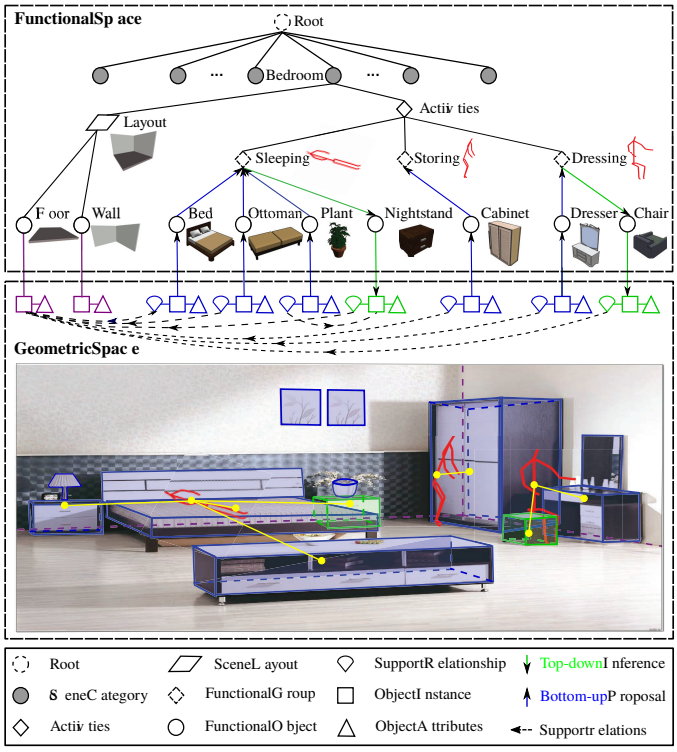


Figure 23: Task-centered representation of an indoor scene [36]. The functional space characterizes the hierarchical structure, and the geometric space encodes the spatial entities with contextual relations. The objects are grouped by the hidden activity groups, *i.e.*, by latent human context.

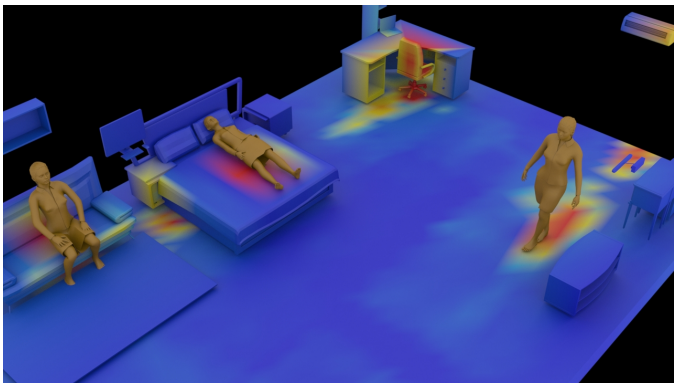


Figure 24: An example of the synthesized human-centric indoor scene (bedroom) with affordance heatmap generated by [95, 266]. The joint sampling of a scene is achieved by the alternative sampling of humans and objects according to the joint probability distribution.

(a) Demonstration (b) Failure by pure imitation

Figure 25: (a) Given a successful human demonstration, (b) the robot may fail to accomplish the same task by imitating the human demonstration due to different embodiment. In this case, a two-finger gripper cannot firmly hold a hammer while swinging; the hammer slips, and the execution fails.

with a Holistic Scene Grammar (HSG). The HSG represents a functional task-centered representation of scenes. As shown in Figure 23, the functional descriptor was composed of functional scene categories, task-centered activity groups, and individual objects. To reverse the scene parsing process with human context, the functionality of scenes can also be adopted in synthesizing new scenes with the human-like arrangement. Qi *et al*. [95, 266] proposed human-centric representations to synthesize the 3D scenes with a simulation engine. As illustrated in Figure 24, they integrate human activities and functional grouping/supporting relations to sample more natural and reasonable activity spaces. In all these prior works, the proposed methods imagine the invisible and latent human poses to help parse and understand the visible scene.

5.3. Functional Manipulation in Robotics

Unlike causality and physics, functionality and affordance are more difficult to evaluate their plausibility. One effective way is to examine whether such information could endow more task capabilities to a system, *e.g.*, a robot. However, due to the difference in morphology between humans and robots, the same object or the same environment does not necessarily introduce the same functionality and affordance. For example, the human hand has five fingers whereas robot gripper usually only has two or three fingers; while a person can firmly grasp a hammer and swing it, a robot might fail as shown in Figure 25. This common problem is known as the "correspondence problem" [267] in Learing from Demonstration (LfD); see two surveys [268, 269]:

To address this issue, the majority of work in LfD usually handcrafts a one-to-one mapping between the human demonstration and the robot execution, restricting the LfD only to mimic the demonstrator's (human's) low-level motor controls and replicate the (almost) identical procedure to accomplish a task. Therefore, the acquired skills can hardly be adapted to new robots or new situations, thereby demanding more robust solutions.

We argue that more explicit modeling knowledge about physical objects and forces is required as we believe the key in imitating manipulation is achieving a *functionally equivalent* manipulation—to imitate and replicate the task execution to achieve the same goal by reasoning about contact forces, instead of merely replicating the trajectory.

However, measuring human manipulation forces is difficult due to the lack of proper, accurate instruments and constraints imposed by measurement devices to natural hand motions. For example, a vision-based manipulation force sensing method [215] often has limitations in handling self-occlusions and occlusions caused during manipulations. Other force sensing devices such as strain gauge FlexForce [270] or the liquid-metal embedded elastomer sensor [271] can be adopted as glove-based systems; but they can be too rigid to conform to the contours of the hand, resulting in limitations on natural hand motion during fine manipulative actions. Recently, Liu *et al*. [272] introduces Velostat, a soft piezoresistive conductive film whose resistance changes under pressure, to an IMU-based pose sensing glove to reliably record manipulation demonstrations with fine-grained force information. This kind of demon-

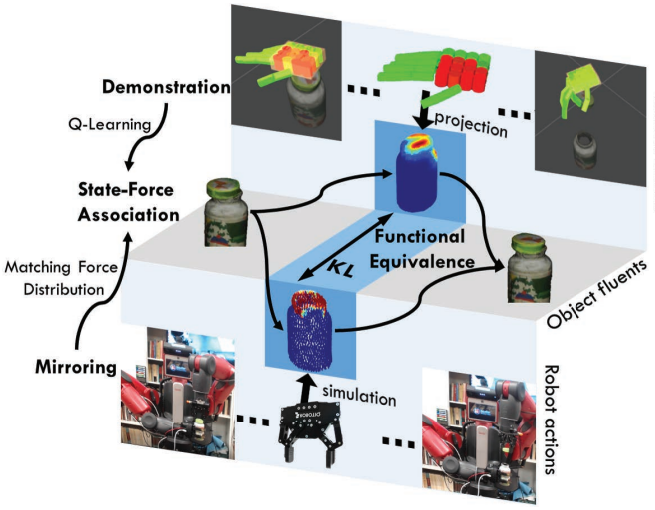


Figure 26: A robot mirrors human demonstrations with functional equivalence [275] by inferring the action that produces similar force, resulting in similar changes of the physical states. Q-Learning is applied to associate types of forces with the categories of the object state changes to produce human-object-interaction (*hoi*) units.

stration is particularly important for the tasks with visually latent changes.

Consider the task of opening medicine bottles that have child-safety locking mechanisms. These bottles require the user to push or squeeze in various places to unlock the cap. By design, attempts to open these bottles using a standard procedure will result in failure. Even if the agent visually observes a successful demonstration, direct imitation of this procedure will likely omit critical steps in the procedure; the visual procedure for opening both medicine and traditional bottles are typically identical. By adopting the glove with Velostat [272], the forces imposed to unlock the child-safety mechanisms of medicine bottles become observable. From these observations with latent force, Edmonds *et al*. [273] learn an action planner through both a top-down stochastic grammar model to represent the compositional nature of the task sequence and a bottom-up discriminative model from the observed poses and forces. These two terms are combined during planning to select the next optimal action. An Augmented Reality (AR) interface is also developed on top of this work to allow easy patching of the robot knowledge [274].

However, the above work is still limited in the sense that the robot actions are pre-defined and the underlying structure of the task is not modeled. Recently, Liu *et al*. [275] proposes a *mirroring* approach and the concept of *functional manipulation* that extends the current LfD, through the physics-based simulation, to address the correspondence problem; see Figure 26. Rather than over-imitating the motion trajectories from the demonstration, it is advantageous for the robot to seek *functionally equivalent* but possibly visually different actions that can produce the same effect and achieve the same goal as those in the demonstration. In particular, the approach has three characteristics compared to the standard LfD. *Force-based*: Beyond visually observable space, these tactile-enabled demonstrations capture a deeper understanding of the physical world that a robot interacts with, providing an extra dimension to address the correspondence problem. *Goal-oriented*: A "goal" is defined as the desired state of the target object and is encoded in a grammar model. The terminal node of the grammar model is the state changes caused by the forces, independent of the embodiments. *Mirroring without overimitation*: Different from the classic LfD, a robot does not necessarily mimic every action in the human demonstration. Instead, the robot reasons about the action to achieve the goal states based on the learned grammar and the simulated forces.

6. Perceiving Intention - The Sense of Agency

Apart from inanimate physical objects, we live in a world with a plethora of animate, goal-directed, intentional agents, whose agency implies the ability to perceive, plan, decide, and to change the environment. Crucially, it entails (1) *intentionality* [276] to represent the goal-state in the future and equifinal variability [277] to be able to achieve the intended goal-state with different actions in various contexts, and (2) *rationality of actions* in relation to their goal [278] to produce the most efficient action available. Perception and comprehension of intent enables humans to better understand and predict the behavior of other agents and engage in cooperative activities with shared goals and intentions with others. The construct of intention, as a basic organizing principles guiding how we interpret one another, has been increasingly granted a central position within accounts of human cognitive functioning, thus should be an essential component for future AI.

In this section, we start with a brief introduction to what constitutes the concepts of "agency," which are deeply rooted in humans as early as 6 months; see Section 6.1. Next, we explain the *rationality* principle as the mechanism behind how both infants and adults perceive animate objects as intentional beings; see Section 6.2. Intention prediction is related to action prediction in modern computer vision and machine learning, but it is much more than predicting an action label; see Section 6.3 from a philosophical view. In Section 6.4, we conclude this section by providing a brief review on the building blocks for intention in computer vision.

6.1. The Sense of Agency

In literature, Theory of Mind (ToM) refers to the ability to attribute mental states, including beliefs, desires, intentions, *etc*., to oneself and others [279], where perceiving and understanding intention is the ultimate goal based on an agent's *belief* and *desire*, since people act in large part to fulfill intentions arising from their beliefs and desires [280].

Evidence from developmental psychology shows that 6-month-olds see human activities as goal-directed behavior [281]. By the age of 10 months, infants segment continuous behavior streams into units that correspond to what adults would see as separate goal-directed acts rather than mere spatial movements or muscle movements [282, 283]. After their first birthdays, infants begin to understand that an actor may

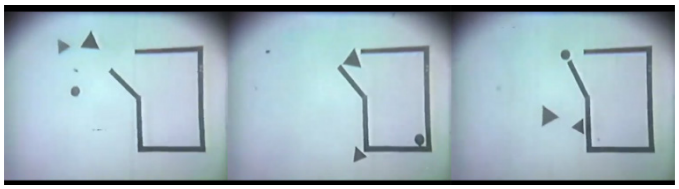


Figure 27: The seminal Heider-Simmel experiment [288]. Adults can perceive and attribute mental states merely from simple geometric shape motions.

consider various action plans to pursue a goal and choose one to enact in intentional action based on reality [284]. 18-month-old infants are able to both *infer* and *imitate* the intended goal of the action even if the action repeatedly fails to achieve the goal [285]. Moreover, infants can evaluate the action's situational constraints and then imitate it in a rational, cost-efficient way, instead of merely copying actions, indicating infants have a deep understanding of relations between the environment, the action, and the underlying intentions [286]. Infants can also recover intentional relations at varying analysis levels, including concrete action goals, higher order plans, and collaborative goals [287].

From infancy onward, we readily process action in intentional terms, despite the complexity of the behavioral stream we actually witness [280]. It is the underlying *intentions*, rather than the surface behaviors, that matter when we observe motions. One latent intention could make several distinctly dissimilar movement patterns cohere conceptually. Even the exact same physical movement could have various different meanings depending on the underlying intentions; *e.g.*, the underlying intention of reaching for a cup could be to fill or clean the cup. Thus, the inference about others' intentions provides the 'gist' of human actions. Research found that humans do not encode the full details of human motion in space; instead, we perceive the motions in terms of intentions—it is the constructed understanding of actions in terms of the actors' goals and intentions that we humans encode in memory and later retrieve [280]. The way to read intentions even creates species-unique forms of cultural learning and cognition [284]. From infants to complex social institutions, the world where we live is constituted from intentions of the agents present [289, 290, 284].

6.2. From Animacy to Rationality

Human vision possesses a unique social nature to extract latent mental states about goals, beliefs, and intents from just visual stimuli. Surprisingly, such visual stimuli do not need to contain rich semantics or visual features for an average human to infer the latent mental states. An iconic illustration is the seminal Heider-Simmel display created in 1940s [288]; see Figure 27. Upon viewing the 2D motion of three simple geometric shapes roaming around, human participants, without any additional hints, automatically and even irresistibly perceive "social agents" with a set of rich mental states, such as goals, emotions, personalities, coalitions, *etc*. These mental states together form a story-like description of the display, such as a hero saving a victim from a bully. Note that in this experiment, where no specific directions or instructions to perceive the objects are provided, participants still tended to describe the objects as being of different sexes and dispositions. Another crucial observation is that human participants always reported the animated being "opens" or "closes" the door, similar to Michotte's "entrance" displace [76]; the movement of the animated being is imparted to the door by the prolonged contact rather than a sudden impact. Such an interpretation of simple geometries as animated beings instead of shapes is a remarkable demonstration of how human vision is able to extract rich social relations and mental states from sparse, symbolized inputs with very minimal visual features.

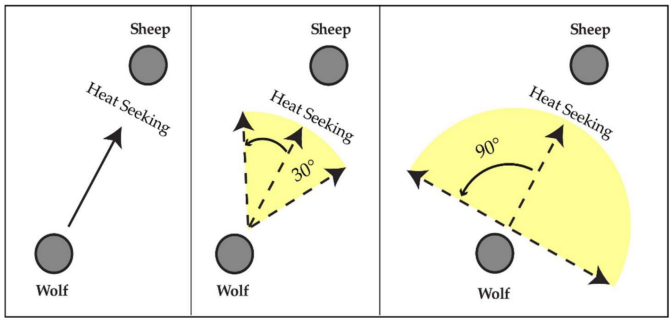


Figure 28: An illustration of the *chasing subtlety* manipulation in the "Don't-Get-Caught" experiment [291]. When the chasing subtlety is 0, the wolf always heads directly toward the (moving) sheep, in a "heat-seeking" manner. When the chasing subtlety is 30, the wolf is always heading in the general direction of the sheep, but is not perfectly heat-seeking: instead, it can move in any direction within a 60 window, with the window always centered on the (moving) sheep. When the chasing subtlety is 90, the wolf's direction of movement is even less constrained: now the wolf may head in an orthogonal direction to the (moving) sheep, but can still never be heading away from it.

In the original display, it is unclear that whether such a visual perception of social relations and mental states was attributed more or less to the dynamic motion of the stimuli or the relative attributes (size, shape, *etc*.) of the protagonists. Berry and Misovich designed a quantitative evaluation for these two confounded variables by degrading the structural display while preserving the original dynamics [293]. They reported a similar number of anthropomorphic terms as in the original design, indicating the structure is not the critical information, which further strengthen the original finding that human perception of the social relations is beyond visual features. Critically, when Berry and Misovich used the static frames, both in the original display and the degraded display, the number of anthropomorphic terms dropped significantly, implying that the dynamic motion and temporal contingency are the crucial factor for the successful perception of the social relations and mental states. This phenomenon was later further studied by Bassili in a series of experiments [294].

Similar simulations of biologically meaningful motion sequences were produced by Dittrich and Lea [295], in which they used simple displays of moving letters. Participants were asked to identify one letter acting as a "wolf" chasing one of the other "sheep" letters, or a "lamb" is trying to catch up with its mother "sheep." Their findings were accorded with the Heider-Simmel experiment—motion dynamics play an important factor in the perception of intentional motion. Specifically, the intentionality

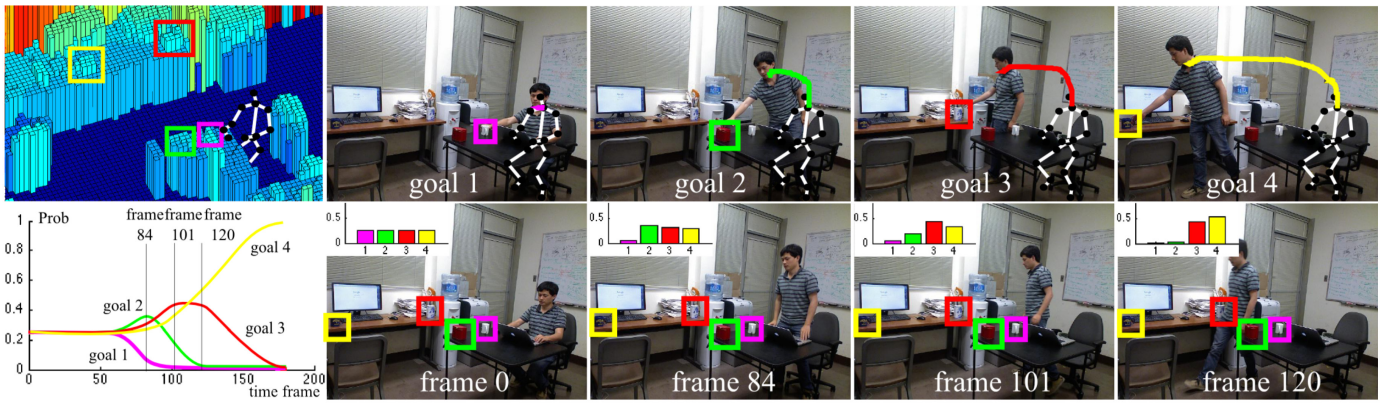


Figure 29: The plan inference task presented in [292]; seen from the perspective of an observing robot. The top left panel shows 4 different goals (target objects) in a 3D scene. The bottom left panel shows one outcome of the proposed method: the marginal probability of each terminal action over time. Note that terminal actions are marginal probabilities over the probability density described by the hierarchical graphical model. The remaining four images on the first row show four rational hierarchical plans for different goals: Goal 1 is within reach, which does not require standing up; Goal 2 requires standing up and reaching out; Goal 3 and Goal 4 require standing up, moving, and reaching for different objects. The second row shows a progression of time corresponding to the bottom left panel. The action sequence and its corresponding probability distributions for each of these four goals are visualized in the bar plots in the upper left of each frame.

appears stronger when the "wolf/lamb" path was more directly related to its target, and it is more salient when the speed difference is significant. Furthermore, they failed to find any significantly different effects when the task was described in neutral (letters) or intentional (*i.e.*, wolf and sheep).

Taking together, these experiments demonstrated even the simplest of moving shapes are irresistibly perceived in an intentional and goal-directed "social" term—a holistic understanding of the events by unfolding the story with goals, beliefs, and intents. A question naturally raises: what is the underlying mechanism for human visual system to perceive and interpret the world with such a rich social context? One possible mechanism governing this process, as proposed by several philosophers and psychologists, is the intuitive agency theory that embodies the so-called "rationality" principle; it states that humans view themselves and others as *causal* agents: (1) devote their *limited* time and resources only to the actions that change the world in accord with their intentions and desires; and (2) achieve their intentions *rationally* by maximizing their *utilities* while minimizing their *costs* given their *beliefs* about the world [296, 278, 297].

Guided by this principle, Gao *et al*. [291] has explored the psychophysics of chasing, one of the most salient and evolutionarily important type of intentional behavior. In an interactive "Don't-Get-Caught" game, a human participant pretended to be a sheep. The task is to detect a hidden "wolf" and keep away from it for 20 seconds. The effectiveness of the wolf's chasing is measured by the percentage of human's failed escapes from it. Across trials, the wolf's pursuit strategy is manipulated by a variable called *chasing subtlety*, which controls the maximum division from the perfect heat-seeking direction; see Figure 28. The results show that human can effectively detect and avoid wolf with small subtlety values, and the wolf with modern subtlety values turns out to be the most "dangerous"—they can still effectively approach the sheep overtime, and the deviation from the most efficient heat-seeking direction severely disrupts human perception of chasing, making themselves undetected; in other words, they can effectively stalk the human-controlled sheep without being noticed. This result is consistent with the "rationality principle" that human perception assumes an agent's intentional action to maximize the efficiency.

Not only are adults sensitive to the cost of actions as demonstrated above, six- to twelve-month-old infants have also shown similar behavior measured in terms of habituation; they tend to look longer when an agent takes a long circuitous route to a goal when a shorter route was available [298, 299]. Crucially, they interpret actions as directed toward goal objects, looking longer when an agent reaches to a new object, even if the reach follows a familiar path [281]. Recently, Liu *et al*. [297] performed five looking-time experiments with 3-month-old infants, in which infants viewed object-directed reaches that varied in efficiencies (following the shortest physically possible path vs. a longer path), goals (lifting an object vs. causing a change in its state), and causal structures (action on contact vs. action at a distance and after a delay). Their experiments verified that infants interpret actions they cannot yet perform as causally efficacious: when people reach for and cause state changes in objects, young infants interpret these actions as goal-directed and look longer when they are inefficient rather than efficient. Such an early-emerging sensitivity to the causal powers of agents engaged in costly and goal-directed actions may provide one important foundation for the rich causal and social learning that characterizes our species.

The rationality principle has been formally modeled as inverse planning governed by Bayesian inference [100, 300, 110]. Planning is a process by which intention causes action. Inverse planning, by inverting the rational planning model via Bayesian inference that integrates the likelihood of observed actions with the prior of mental states, can infer the latent mental intentions. Based on inverse planning, Baker *et al*. [100] proposed a framework for goal inference, where the bottom-up information of behavior observations and the top-down prior knowledge of goal space are integrated to allow the inference

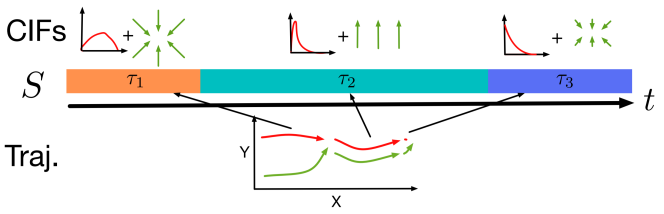


Figure 30: Inference of human interaction from motion trajectories [108]. The top demonstrates the change of conditional interactive field (CIF) in sub-interactions as the interaction proceeds, where the CIF models the expected relative motion pattern conditioned on the reference agent's motion. The bottom indicates the change of interactive behaviors in terms of motion trajectories. The colored bars in the middle depict the types of the sub-interactions.

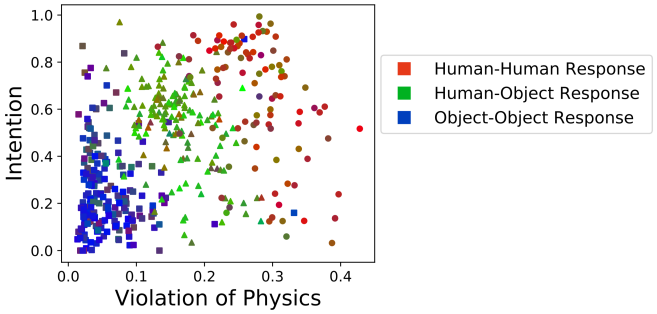


Figure 31: Constructed psychological space including HH animations with 100% animacy degree, HO animations, and OO animations. Here, a stimulus is depicted by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. The two coordinates of the space are the averaged measures between the two entities, as the measure of the degree of violation of physical laws (horizontal) and the measure of values indicating the presence of intention. The mark shapes of data points correspond to the interaction types used in the simulation for generating the corresponding stimuli (circle: HH, triangle: HO, square: OO).

of the underlying intention. In addition, Bayesian networks, with its flexibility for representing probabilistic dependencies and causal relations, as well as the efficiency of inference methods, have proven to be one of the most powerful and successful approaches for intention recognition [301, 302, 303, 300].

Going from the symbolic input to real video input, Holtzen *et al*. [292] presented an inverse planning method to infer human hierarchical intents from partially observed RGB-D videos; their algorithm is able to infer human intents by reverse-engineering the decision making and action planning processes in human minds under a Bayesian probabilistic programming framework; see Figure 29. The intents are represented as a novel hierarchical, compositional, and probabilistic graph structure, describing relationships between actions and plans.

By bridging the abstract Heider-Simmel animations and aerial videos, Shu *et al*. [108] proposed a method to infer humans' intention to interact from motion trajectories; see Figure 30. A non-parametric exponential potential function is learnt to derive the "social force and fields" by calculus of variations (as in Landau physics); such force and field explain human motion and interactions in the collected drone videos. The model provides a good fit to human judgments of interactiveness and is able to synthesize decontextualized animations with a controlled degree of interactiveness.

In outdoor scenarios, Xie *et al*. [69] jointly infer object functionality and human intentions by reasoning human activities. Based on the "rationality" principle, people in the observed videos are expected to intentionally take shortest paths towards functional objects subject to obstacles, where people can satisfy certain needs (*e.g*., a vending machine can quench thirst); see Figure 7a. Here, the functional objects are "dark matter" since they are typically hard to detect in low-resolution surveillance videos and have the functionality to "attract" people. Xie *et al*. formulate the agent-based Lagrangian mechanics wherein human trajectories are probabilistically modeled as motions in many layers of "dark energy" fields, where each agent can select a particular force field to affect its motions, thus define the minimum-energy Dijkstra path toward the corresponding source "dark matter." Such a model is effective in predicting human intentional behaviors and trajectories, localizing functional objects, and discovering distinct functional classes of objects by clustering human motion behavior in the vicinity of functional objects.

6.3. *Beyond Action Prediction*

Intention is related to action prediction in modern computer vision [305], *much* more than predicting merely an action label; humans have a strong and early inclination to interpret actions in terms of intention as long-term *social learning* of novel means and novel goals. From a philosophical view, Csibra *et al*. [99] contrasted three distinct mechanisms: (1) action-effect association, (2) simulation procedures, and (3) teleological reasoning. They concluded that action-effect association and simulation can only serve action monitoring and prediction; social learning, in contrast, requires the inferential productivity of teleological reasoning.

Simulation theory claims that the mechanism underlying the attribution of intentions to actions might rely on simulating the observed action and mapping it onto our own experiences and intention representations [306], and such simulation processes are at the heart of the development of intentional action interpretation [285]. In order to understand others' intention, humans subconsciously empathize with the person they are observing, estimate what their own actions and intentions might be in that situation. Here, action-effect association [307] plays an important role in quick online intention prediction, and the ability to encode and remember these two component associations contributes to infants' imitation skills and intentional action understanding [308]. Accumulating neurophysiological evidences support such simulations in human brain, such as the mirror neurons [309], which has been linked to intention understanding by many studies [310, 97]. However, some studies also find that infants are capable of processing goal-directed actions before they have the ability to perform the actions themselves(*e.g*., [311]), which poses challenges to the simulation theory.

To address social learning, teleological action interpretational system [312] takes a 'functional stance' for the computational representation of goal-directed action [99], where

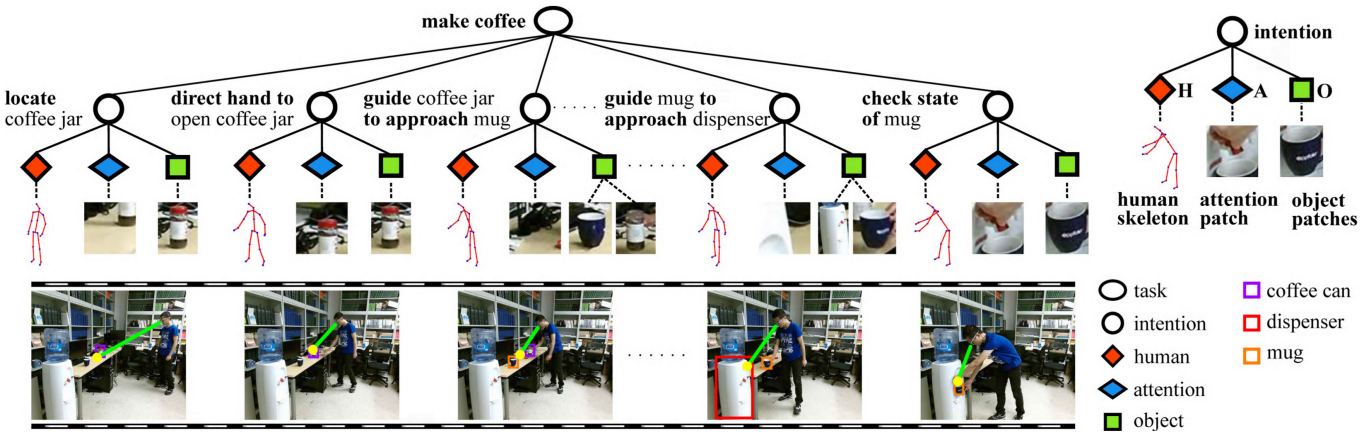


Figure 32: Model task as sequential intentions by Human-Attention-Object (HAO) graph [304].

such teleological representations are generated by the aforementioned inferential "rationality principle" [313]. In fact, the very notion of 'action' implies a motor behaviour performed by an agent, which is conceived in relation to the end state it is destined to achieve. Attributing a goal to the observed action enables humans to predict the future course, to evaluate the causal efficacy, and to justify the action. Also, action predictions can be made by breaking down the path towards the goal into sub-goals in a hierarchical fashion, eventually arriving at elementary motor acts, such as grasping.

These three mechanisms do not compete but complement each other. The fast effect prediction provided by action-effect associations can serve as a starting hypothesis for teleological reasoning or simulation procedure; the solutions provided by teleological reasoning in social learning can also be stored as action-effect associations for subsequent rapid recall.

6.4. Building Blocks for Intention in Computer Vision

Understanding and predicting human intentions from images and videos is a research topic driven by many real-world applications, including visual surveillance, human-robot interaction, autonomous driving vehicle, *etc*. In order to better predict intention based on pixel inputs, it is necessary and indispensable to fully exploit comprehensive cues, such as motion trajectory, gaze dynamics, body posture and movements, human object relations, and communicative gestures (*e.g*., pointing).

Motion trajectory alone could be a strong cue for intention prediction as discussed in Section 6.2. With intuitive physics and perceived intention, humans also demonstrate the ability to distinguish social events from physical events with very limited motion trajectory stimuli, *e.g*., movements of a few simple geometric shapes. Shu *et al*. studied the underlying computational mechanisms and proposed a unified psychological space that reveals the partition between the perception of physical events involving inanimate objects and the perception of social events involving human interactions with other agents [109]. This unified space consists of two prominent dimensions: (1) an intuitive sense of whether physical laws are obeyed or violated; and (2) an impression of whether an agent possesses intentions as inferred from movements of simple shapes; see Figure 31. Their experiments demonstrate that the constructed psychological space successfully partitions human perception of physical versus social events.

Eye gaze also plays an important role in reading other peoples' minds, being closely related to the underlying attention, intention, emotion, personality, and tied to what human is thinking and doing [314]. Evidence from psychology suggests that eyes are a cognitively special stimulus, with unique "hard-wired" pathways in the brain dedicated to their interpretation; humans have the unique ability to infer others' intentions from eye gazes [315]. Social eye gaze functions also transcend cultural differences, forming a kind of universal language [316]. Computer vision systems heavily rely on gazes as cue for intention prediction from real scene images and videos. For instance, Wei *et al*. [304] jointly infers human attention, intentions, and tasks from videos. Given an RGB-D video where a human performs a task, they answer three questions simultaneously: (1) where the human is looking—attention/gaze prediction, (2) why the human is looking—intention prediction, and (3) what task the human is performing—task recognition. They proposed a hierarchical model of human-attention-object (HAO) which represents tasks, intentions, and attention under a unified framework. A task is represented as sequential intentions which transition to each other. An intention is composed of human pose, attention and objects; see Figure 32.

Communicative gazes and gestures (*e.g*., pointing) stand out for intention expression and perception in collaborative interactions. Humans need to recognize partners' communicative intention to successfully collaborate with others and survive in the world. Human communication in mutualistic collaboration often involves agents informing recipients of things they believe will be useful or relevant to them. Tomasello *et al*. [318] investigated whether pairs of chimpanzees were capable of communicating to ensure coordination during collaborative problem-solving. In their experiments, the chimpanzee pairs needed two tools to extract fruits from an apparatus. The communicator in each pair could see the location of the tools (hidden in one of

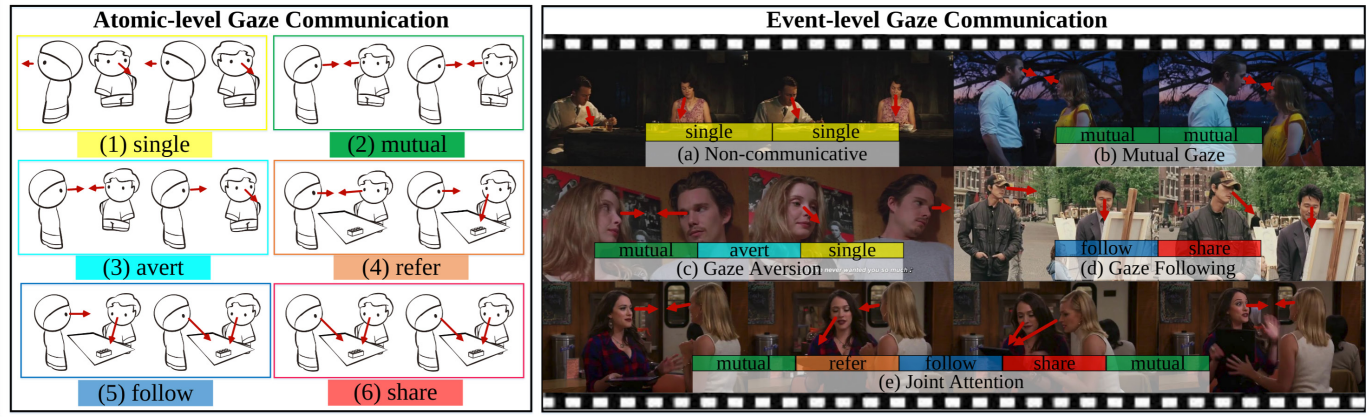


Figure 33: Human gaze communication dynamics in two hierarchical levels [317]: (1) atomic-level gaze communication describes the fine-grained structures in human gaze interactions, and (2) event-level gaze communication refers to long-term social communication events temporally composed of atomic-level gaze communications.

two boxes), whereas only the recipient could open the boxes. The communicator increasingly communicated the tools' location by approaching the baited box and giving the key needed to open it to the recipients. The recipient used these signals and obtained the tools, transferring one of the tools to the communicator so that the pair could collaborate in obtaining the fruits. As demonstrated by this study, even chimpanzees already obtain the necessary socio-cognitive skills to naturally develop a simple communicative strategy to ensure coordination in a collaborative task. For human communicative gaze dynamics, Fan *et al*. [319] studied the problem of inferring shared eye gazes in third-person social scene videos, which is a phenomenon that two or more individuals simultaneously look at a common target in social scenes. A follow-up work [317] studied all kinds of gaze communications in social activities from both atomic-level and event-level; see Figure 33. A spatio-temporal graph network is proposed to explicitly represent the diverse interactions in the social scenes and infer atomic-level gaze communications by message passing.

Humans communicate intentions multimodally, thus facial expression, head pose, body posture and orientation, arm motion, gesture, proxemics, and relations with other agents and objects can all contribute to human intention analysis and comprehension. Researchers in robotics try to incorporate with robots such ability to act naturally and properly subject to "social affordance", which represents action possibilities following basic social norms. Trick *et al*. [320] propose an approach for multimodal intention recognition considering four modalities, including speech, gestures, gaze directions, and scene objects, focusing on uncertainty reduction through classifier fusion. Shu *et al*. [321] presents a generative model for robot learning of social affordance from human activity videos. By discovering critical steps (*i.e*., latent sub-goals) in an interaction and learning structural representations of human-human and human-object-human interactions, describing how agents' body-parts move and what spatial relations they should maintain to complete each sub-goal, robots can infer its own movement in reaction to the human body motion. Such a social affordance could also be represented by a hierarchical grammar model [322], enabling a real-time motion inference for human-robot interaction; the learned model was demonstrated to successfully infer human intention and generate human-like socially appropriate responding behaviors for robots.

7. Discussions and Summary

Below, we discuss a few topics related to functionality, physics, intention, and causality (FPIC) that have been less explored in the field of computer vision. We hope to shed some lights on these potential directions for future FPIC research.

- **Causality:** Studying causes and its effects opens a new direction of analogy and relational reasoning [324]. Apart from the four-term analogy, or proportional analogy, John C. Raven [325] proposed the Raven's Progressive Matrices Test (RPM) in the image domain. The RAVEN dataset [323] is introduced in the computer vision community and serves as a systematic benchmark for various visual reasoning models. Empirical studies show that abstract-level reasoning combined with effective feature extraction models could notably improve the performance of reasoning, analogy, and generalization. However, the performance gap between human and computational models calls for future research into this field; see Section 7.1.
- **Physics:** Physics-based simulation for multi-material multi-physics phenomena (see Figure 35) will continue to play an important role physical scene understanding, functional reasoning, and many related AI problems. We argue that cognitive AI needs to accelerate the pace of adopting more advanced simulation models from computer graphics, in order to benefit from the capability of highly predictive forward simulations, especially their GPU optimization which allows for real-time performance [326]. Here, we provide a brief review of the recent physics-based simulation methods, in particular, Material Point Method (MPM); see Section 7.2.
- **Functionality:** As mentioned in Section 5.3, it is difficult to evaluate the plausibility of the functionality and affordance

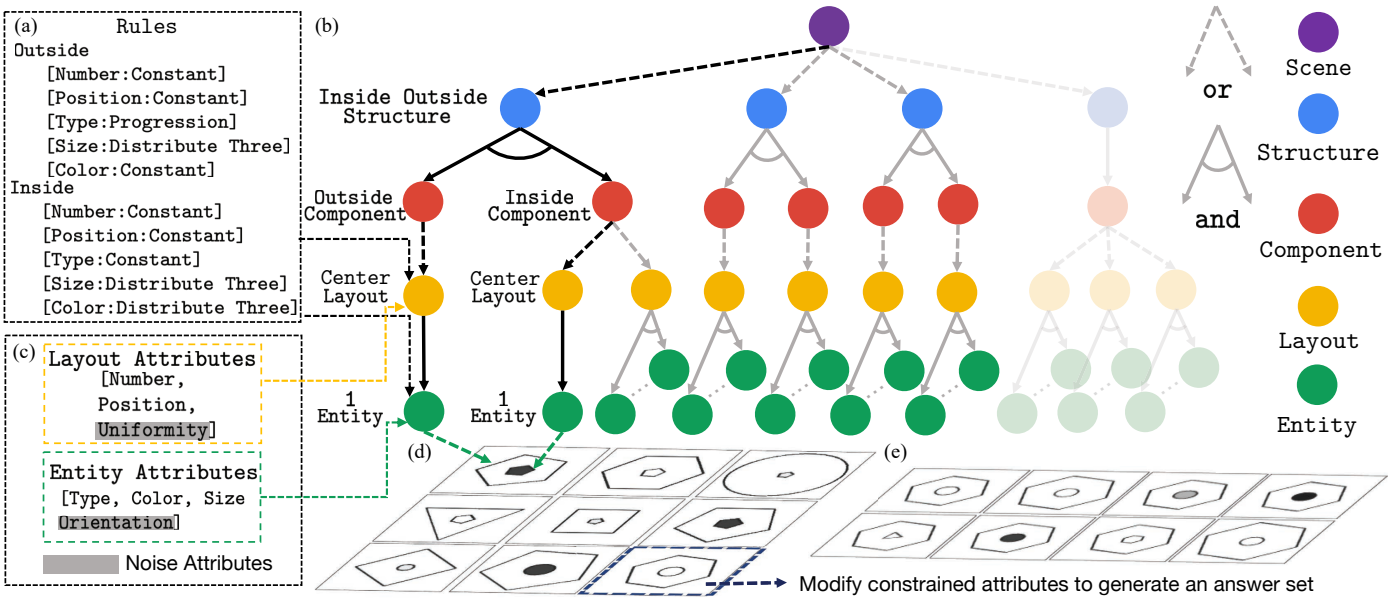


Figure 34: RAVEN creation process proposed in [323]. A graphical illustration of (a) the grammar production rules used in (b) Attributed Stochastic Image Grammar (A-SIG). (c) Note that `Layout` and `Entity` have associated attributes. (d) A sample problem matrix and (e) a sample candidate set.

inferred in a 3D scene. Here, we argue that the ultimate standard for validating the functionality and affordance is to examine whether an agent is capable of (1) accomplishing the very same task using different sets of objects with different orders and/or sequences of actions in different environments, and (2) rapidly adapting such learned knowledge to new tasks. By leveraging the state-of-the-art game engines, we begin to explore this possibility in a large scale; see Section 7.3.

- **Intention:** Given an image or a video in which agents interacting with a 3D scene, we can mostly assume that the observed person makes near optimal choices to minimize the cost of certain tasks; *i.e.*, no deception or pretense. This is known as rational choice theory; *i.e.*, a rational person's behavior and decision-making are driven by maximizing its utility function. In the field of mechanism design in economics and game theory, this is known as revelation principle, in which we assume each agent *truthfully* report their preference; see [327] for a short and introductory survey. Building computational models for human utilities could be traced back to the English philosopher, Jeremy Bentham, and his works on ethics known as utilitarianism [328]. We argue such a utility-driven learning could be more invariant than the traditional supervised training for computer vision and AI; see Section 7.4.

7.1. From Causality to Analogy and IQ

In psychology literature, we call two cases analogous if they share a common *relationship*. Such a relationship does not need to be within the same category in terms of the same label commonly adopted in computer vision; rather, it emphasizes the commonality on a more abstract-level. For instance, according to [329], the earliest major scientific discovery with analogy can be dated back to the era of imperial Rome, when they made an analogy between the sound saves and the water waves, sharing a few similar patterns; *e.g.*, the intensity will diminish when propagating across the space. The key to such a successful analogy is to understand *causes and their effects* [330].

The history of analogy can be categorized into three streams of research; see [324] for a capsule history and review of the literature. One stream lies in the psychometric tradition as four-term or "proportional" analogies; the earliest discussions can be traced back to Aristotle [331]. An example in AI is the *word2vec* model [332, 333], capable of making the four-term word analogy, *e.g.*, [king:queen::man:woman]. In the image domain, a similar test was invented by John C. Raven [325]—the RPM.

RPM has been widely accepted and believed to be highly correlated with real intelligence [334]. Unlike Visual Question Answering (VQA) [335] in computer vision at the periphery of the cognitive ability test circle [334], RPM lies directly at the center of human intelligence, is diagnostic of abstract and structural reasoning ability [336], and characterizes the defining feature of high-level intelligence, *i.e.*, *fluid intelligence* [337]. It has been shown that RPM is harder than existing visual reasoning tests in the following ways [323].

- Unlike VQA where natural language questions usually imply what to pay attention to in the image, RPM relies merely on visual clues provided in the matrix and the *correspondence problem* itself, *i.e.*, finding the corresponding objects across frames for relation extraction, is already a major factor distinguishing populations of different intelligence [334].
- While current visual reasoning tests only require spatial and semantic understanding, RPM needs joint spatial-temporal reasoning in the problem matrix and the answer set. The limit of *short-term memory*, the ability of *analogy*, and the discovery of the *structure* have to be taken into consideration to solve a RPM problem.

- Structures in RPM make the compositions of rules much more complicated. Problems in RPM usually include more sophisticated logic with recursions. Combinatorial rules composed at various levels also make the reasoning progress extremely difficult.

To push the limit of current vision systems' reasoning and analogy-making ability, the Relational and Analogical Visual rEasoNing dataset (RAVEN) [323] was created to promote further research in this area. The dataset is designed inherently focus on reasoning and analogy-making instead of visual recognition. It is unique in the sense that it builds a semantic link between visual reasoning and structure reasoning in RPM by grounding each problem into a sentence derived from an Attributed Stochastic Image Grammar (A-SIG): each instance is a sentence sampled from a pre-defined A-SIG and a rendering engine transforms the sentence to the corresponding image. See Figure 34 for a graphical illustration of the generation process. This semantic link between vision and structure representation opens new possibilities by breaking down the problem into image understanding and abstract-level structure reasoning. Zhang *et al*. [323] empirically demonstrated that models with a simple structure reasoning module to incorporate both vision-level understanding and abstract-level reasoning and analogy-making would notably improve their performance in RPM, while various previous approaches on relational learning perform only slightly better than random guess.

Analogy consists of more than mere spatiotemporal parsing and structural reasoning. For example, *contrast effect* [338] is proven to be one of the key ingredients in relational and analogical reasoning for both human and machine learning [339, 340, 341, 342, 343]. Originated from perceptual learning [344, 345], it is well established in the field of psychology and education [346, 347, 348, 349, 350] that teaching new concepts by comparing with noisy examples is quite effective. Smith and Gentner [351] summarize that comparing cases facilitates transfer learning and problem-solving, as well as the ability to learn relational categories. Gentner [352] in his structure-mapping theory postulate that learners generate a structure alignment between two representation when they compare two cases. A later article [353] firmly supports this conjecture and shows finding the individual difference is easier for humans when similar items are compared. A more recent study from Schwartz *et al*. [354] also shows that contrasting cases help foster an appreciation of a deep understanding of concepts.

To retrieve this missing treatment of contrast in machine learning, computer vision, and more broadly in AI, Zhang *et al*. [355] proposes a learning perceptual inference by contrast that explicitly introduces the notion of contrast in model training. Specifically, a contrast module and a contrast loss are incorporated into the algorithm at the model level and the objective level, respectively. The permutation-invariant contrast module summarizes the common features from different objects and distinguishes each candidate by projecting it onto its residual on the common feature space. The final model that comprises ideas from contrast effects and perceptual inference achieves the state-of-the-art performance on major RPM datasets.

7.2. A Brief Review of Recent Physics-based Simulation

The accuracy of physics-based reasoning greatly relies on the fidelity of the physics-based simulation. Similarly, the scope of supported virtual materials and their physical interactions directly determines the complexity of the corresponding AI tasks. In computer graphics, many mathematical and physical models have been developed, and have been applied to simulation of various solids and fluids in a 3D virtual environment, since the pioneering work of Terzopoulos *et al*. [356, 357] for solids and Foster *et al*. [358] for fluids.

For decades, the computer graphics and computational physics community seek to increase the robustness, efficiency, stability, and accuracy for simulating cloth, collisions, deformables, fire, fluids, fracture, hair, rigid bodies, rods, shells, and many other substances. Computer simulation-based engineering science plays an important role in solving various modern problems as an inexpensive, safe, and analyzable companion to physical experiments. The most challenging problems are those involving extreme deformation, topology change, and interactions among various materials and phases. Examples of these problems include hyper-velocity impact, explosion, crack evolution, fluid-structure interactions, climate simulation, and ice-sheet movements, *etc*. Despite the rapid development of computational solid and fluid mechanics, effectively and efficiently simulating these complex phenomena remains difficult. Based on how to discretize the continuous physical equations, existing methods can be classified into:

1. Eulerian grid-based approaches, where the computational grid is fixed in space, and physical properties advect through the deformation flow. A typical example is the Eulerian simulation of free surface incompressible flow [359, 360]. Eulerian methods are more error-prone and require delicate treatment in dealing with deforming material interfaces and boundary conditions since no explicit tracking of them is available.
2. Lagrangian mesh-based methods, represented by FEM [361, 362, 363], where the material is described with and embedded in a deforming mesh. Mass, momentum, and energy conservation can be solved with less effort. The main problem of FEM is the mesh distortion and the lack of contact during large deformation [364, 365] or topologically changing events [366].
3. Lagrangian mesh-free methods, such as Smoothed Particle Hydrodynamics (SPH) [367] and Reproducing Kernel Particle Method (RKPM) [368]. These methods allow arbitrary deformation but require expensive operations such as neighborhood search [369]. Since the interpolation kernel is approximated with neighboring particles, they also tend to suffer from numerical instability issues.
4. Hybrid Lagrangian-Eulerian methods, *e.g*., Arbitrary Eulerian-Lagrangian Methods (ALE) [370] and Material Point Method (MPM). These methods (particularly MPM) combine advantages of both Lagrangian methods and Eulerian grid methods by using a mixed representation.

In particular, as a generalization of the hybrid Fluid Implicit Particle (FLIP) method [372, 373] from computational

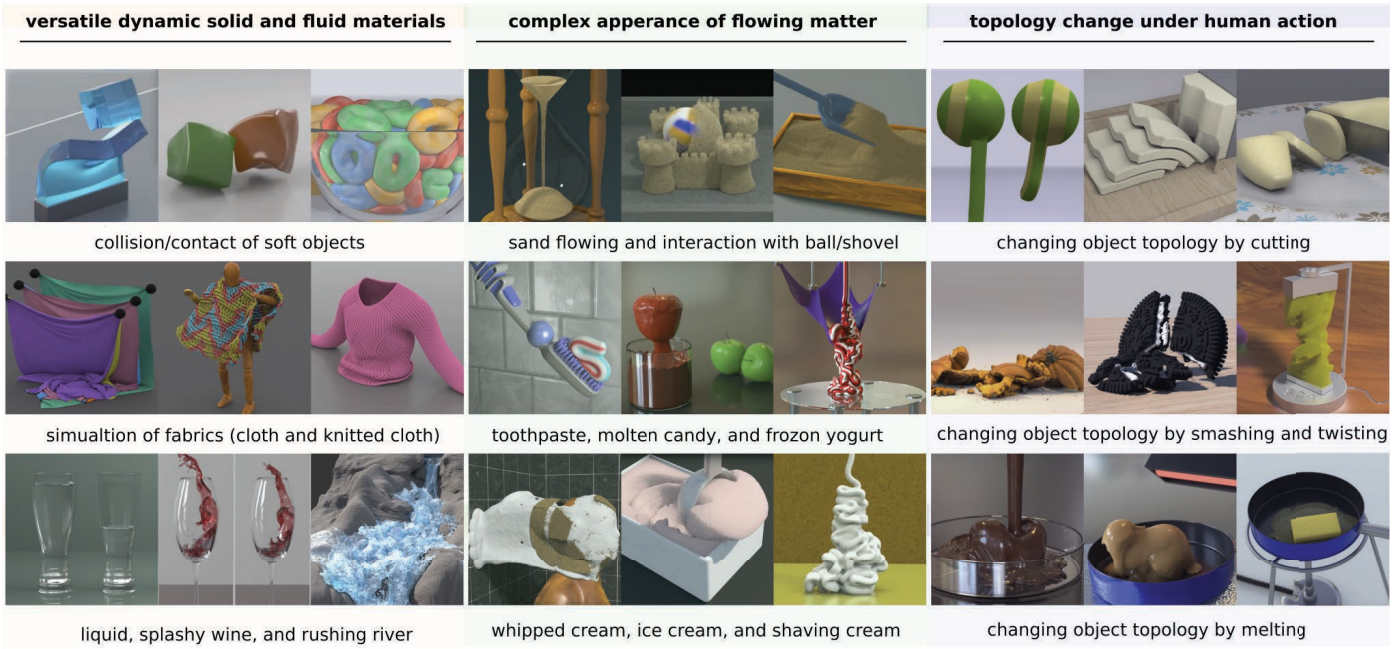
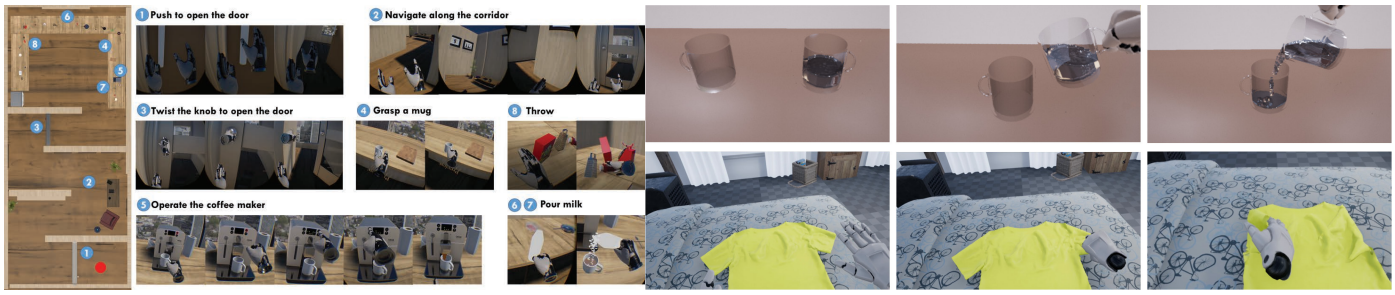


Figure 35: Diverse physical phenomena simulated using MPM.



(a) Various task executions in VRGym

(b) Real-time fluid and cloth simulations in VRGym

Figure 36: VRGym [371], an example of the simulated virtual environment as a large task platform. (a) Inside this platform, either a human agent or a virtual agent can perform various actions in a virtual scene to evaluate the success of a particular task execution. (b) In addition to the rigid body simulation, VRGym also has realistic real-time fluid and cloth simulation by leveraging the state-of-the-art game engines.

fluid dynamics to computational solid mechanics, MPM has proven to be a promising discretization choice for simulating many solid and fluid materials since its introduction two decades ago [374, 375]. In the field of visual computing, the existing work include snow [376, 377], foam [378, 379, 380], sand [381, 382], rigid body [383], fracture [384, 385], cloth [386], hair [387], water [388], and solid-fluid mixtures [389, 390, 391]. In computational engineering science, it has also become one of the most recent and advanced discretization choices for various applications. Due to its many advantages, it has been successfully applied to tackling extreme deformation events such as fracture evolution [392], material failure [393, 394], hyper-velocity impact [395, 396], explosion [397], fluid-structure interaction [398, 399], biomechanics [400], geomechanics [401], and many other examples that are considerably much more difficult with traditional non-hybrid approaches. Besides experiencing the tremendously expanding scope of applications, MPM has also been extensively improved on its discretization scheme [402]. To alleviate numerical inaccuracy and stability issues associated with the original MPM formulation, researchers have proposed different variations of MPM, including Generalized Interpolation Material Point (GIMP) method [403, 404], Convected Particle Domain Interpolation (CPDI) method [405] and Dual Domain Material Point (DDMP) [406].

7.3. Virtual Environments as Task Rich Platforms

A hallmark of machine intelligence is the capability to rapidly adapt to new tasks and "achieve goals in a wide range of environments" [407]. To reach this goal, in recent years, we have seen the increasing use of synthetic data and simulation platforms for indoor scenarios by leveraging the state-of-the-art game engines and publicly available free 3D contents [408, 409, 266, 410], including MINOR [411], HoME [412], Gibson [413], House3D [414], AI-THOR [415], VirtualHome [416], VRGym [371] (see Figure 36), VRKitchen [417], *etc*. Similarly, AirSim [418] was developed for

outdoor scenarios. Such synthetic data could be relatively easily scaled up compared to the traditional data collection and labeling process. With an increasing realism and faster speed of the rendering methods using dedicated hardware, the synthetic data from the virtual world is getting closer ever to the data collected from the physical world.

7.4. From Rationality to Utility Learning

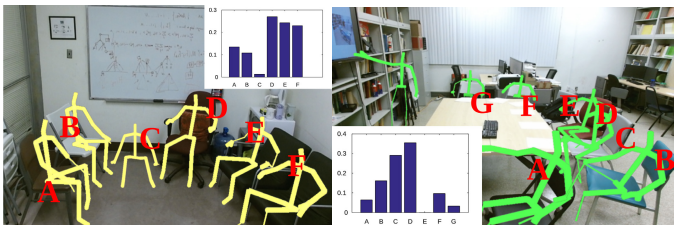
By observing a rational person's behavior and choices, one can reverse-engineer their reasoning and learning process, and estimate their values. Utilities, or values, are also used in the field of AI in planning schemes like Markov decision process (MDP), and are often associated with states of a task. However, in the literature of MDP, the "value" is not a reflection of true human preference and, inconveniently, is tightly dependent on the agent's actions [419]. Here, we briefly review two case studies in computer vision and robotics using a utility-driven learning approach.

- As shown in Figure 37, by observing the choices people make in videos (particularly in selecting a chair in which to sit), a computer vision system [211] is able to learn the comfort intervals of the forces exerted on body parts (while sitting), which accounts for people's preferences in terms of human *internal* utilities.
- Similarly, Shukla *et al*. [420] adopted the idea of learning human utility in order to learn from human demonstrations for a robotics task. A proof of concept work shows a pipeline in which the agent learns the *external* utility of human and plans with the learned utility functions for a cloth-folding task. Specifically, under the assumption that the utility of goal states is better than initial states, this work learns the *external* utility of human by ranking the pairs of states extracted from images.

Such a rational principle has also been studied in the field of linguistic and philosophy, notably from Grice's influence work on the theory of implicature [421]. The core insight of Grice's work was that language use is a form of rational action; thus, technical tools for reasoning about rational action should elucidate linguistic phenomena [422]. Such a goal-directed view of language production has led to a few interesting language games [424, 425, 426, 427, 428, 429], the development of engineering systems for natural language generation [430], and a vocabulary for formal descriptions of pragmatic phenomena in the field of game-theory [431, 432]. More recently, by assuming the communications between the agents to be helpful yet parsimonious, the 'Rational Speech Act' [423, 422] model has demonstrated promising results in solving some challenging referential games.

7.5. Summary

Robots are mechanically capable of performing a wide range of human activities; but in practice, they do very little for us. Robots still lack physical and social common senses, and this limitation inhibits their capacity to aid in our daily lives. In this article, we reviewed four crucial aspects as the building blocks of commonsense: causality, physics, functionality, and intentions. We argue these cognitive aspects are the foundation for constructing a cognitive architecture towards future artificial general intelligence.



(a) (b)

Figure 37: Examples of sitting activities in (a) an office and (b) a meeting room [211]. In addition to geometry and appearance, people also consider other important factors including comfortability, reaching cost, and social goals when choosing a chair. The histograms indicate human preferences for different candidate chairs.

The position and opinions provided in this article do not intend to serve as *the* solution for the future of cognitive AI or artificial general intelligence; rather, we are calling for attention in this rapid developing community to look for emerging and less explored directions by identifying a few crucial aspects that have shown potential to build cognitive AI. In fact, there are many other topics that, we believe, are also the essential AI ingredients; for example:

- **Communication in Multiple Agent Systems (MAS):** Being able to communicate and collaborate with other agents including human beings is a prerequisite of achieving artificial general intelligence. Modeling multiagent communication dates back to 1998, when Kinney *et al*. [433] proposed adaptive learning of multiagent communication strategy as a predefined rule-based control system. To scale up from rule-based systems, decentralized partially observable Markov decision processes were used to model multiagent interaction with communication as a special type of action among agents [434, 435]. As the success of RL in single agent games [436], generalizing Q-learning [437, 438] and actor-critic [439, 440] based methods originating from single agent games to MAS has been a booming topic in recent years. In addition, human dialogues, as one of the most important communication tools, have been widely explored in MAS [441, 442]; reasoning human dialogues with causal relations and Theory of Mind can help machines to generate reasonable human-like responses. Unlike most previous approaches that takes advantage of attention mechanisms to generate context related dialogues, Zheng *et al*. [443] demonstrated an approach to model relations among dialogue turns with explicit graph structures, showing that such structures would not only present reasoning capabilities but also improve their performance in dialogue generation tasks.
- **Formation of Language:** The emergence of language in MAS is also a productive topic in multiagent decentralized collaborations. Modeling communication as a special type of action, and training alone interactions [444, 445, 446], agents can learn how to communicate with continuous signals only decipherable within a group. Emergence of more

realistic communication protocol using discrete messages has been explored with various types of communication games [447, 448, 449, 450], in which agents need to process visual signals like synthesized or natural images and ground discrete tokens to attributes or semantics of the images to form effective protocols. Moreover, by letting groups of agents play communication games spontaneously, several linguistic phenomenon in emergent communication and language have been studied [451, 452, 453].

- **Morality in Social Systems:** Morality is an abstract and complex concept composed of a set of common principles, such as fairness, obligation, and permissibility. It is, in fact, deeply rooted in the trade-offs people make everyday under the guidance of those principles, and the trade-offs often represent innate conflicts posited on the principles centered around morality [454, 455]. Moral judgement is extremely complicated due to the variability in standards among different individuals, social groups, cultures, and even forms of violation to ethical rules. For example, two distinct societies could hold opposite views on preferential treatment of kin: corruption or moral obligation [456]. In addition, the same principle might be viewed differently in two social groups with distinct cultures [457]. Even within the same social group, different individuals might have different standards on the same moral incident or principle [458, 459, 460]. Many works proposed theoretical accounts for categorizing welfare used in morality calculus, including "base goods" and "primary goods" [461, 462], "moral foundations" [463], and ability of value judgment from infants' points of view [464]. Despite its complexity and diversity, morality is an essential piece towards building human-like machines, which requires a computational account of moral judgment. One recent approach combines utility calculus and Bayesian inference to perform moral learning to distinguish and evaluate different moral principles [456, 465, 466].

8. Acknowledgments

This article presents some representative work selected from a US and UK Multidisciplinary University Research Initiative (MURI) collaborative project on visual commonsense reasoning, focusing on human vision and computer vision. The team consists of interdisciplinary researchers in computer vision, psychology, cognitive science, machine learning, and statistics, from both the US (CMU, MIT, Stanford, UCLA, UIUC, and Yale) and the UK (Birmingham, Glasgow, Leads, and Oxford)[2]. This MURI team also holds an annual review meeting at various locations together with two related series of CVPR/CogSci workshops[3][4].

We thank Professor Chenfanfu Jiang at University of Pennsylvania, Dr. Behzad Kamgar-Parsi at Office of Naval Research (ONR), Dr. Bob Madahar at Defence Science and Technology Laboratory (DSTL), Luyao Yuan, Zilong Zheng, Xu Xie, Xiaofeng Gao, and Qingyi Zhao at UCLA, Dr. Yibiao Zhao at ISEE, and Dr. Mark Nitzberg at UC Berkeley for helpful discussions on various sections. This work reported herein is supported by MURI ONR N00014-16-1-2007, DARPA XAI N66001-17-2-4029, and ONR N00014-19-1-2153.

[2] See `https://vcla.stat.ucla.edu/MURI_Visual_CommonSense/` for details about this MURI project.

[3] Workshop on VisionMeetsCognition: Functionality, Physics, Intentionality, and Causality: `https://www.visionmeetscognition.org/`

[4] Workshop on 3D Scene Understanding for Vision, Graphics, and Robotics: `https://scene-understanding.com/`

References

[1] D. Marr, Vision: A computational investigation into the human representation and processing of visual information. MIT Press, Cambridge, Massachusetts, 1982.

[2] M. Mishkin, L. G. Ungerleider, K. A. Macko, Object vision and spatial vision: two cortical pathways, Trends in Neurosciences 6 (1983) 414–417.

[3] M. Land, N. Mennie, J. Rusted, The roles of vision and eye movements in the control of activities of daily living, Perception 28 (11) (1999) 1311–1328.

[4] K. Ikeuchi, M. Hebert, Task-oriented vision, in: Exploratory vision, Springer, 1996, pp. 257–277.

[5] F. Fang, S. He, Cortical responses to invisible objects in the human dorsal and ventral pathways, Nature Neuroscience 8 (10) (2005) 1380.

[6] S. H. Creem-Regehr, J. N. Lee, Neural representations of graspable objects: are tools special?, Cognitive Brain Research 22 (3) (2005) 457–469.

[7] K. Ikeuchi, M. Hebert, Task-oriented vision, in: International Conference on Intelligent Robots and Systems (IROS), 1992.

[8] M. C. Potter, Meaning in visual search, Science 187 (4180) (1975) 965–966.

[9] M. C. Potter, Short-term conceptual memory for pictures, Journal of experimental psychology: human learning and memory 2 (5) (1976) 509.

[10] P. G. Schyns, A. Oliva, From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition, Psychological science 5 (4) (1994) 195–200.

[11] S. Thorpe, D. Fize, C. Marlot, Speed of processing in the human visual system, Nature 381 (6582) (1996) 520.

[12] M. R. Greene, A. Oliva, The briefest of glances: The time course of natural scene understanding, Psychological Science 20 (4) (2009) 464–472.

[13] M. R. Greene, A. Oliva, Recognition of natural scenes from global properties: Seeing the forest without representing the trees, Cognitive Psychology 58 (2) (2009) 137–176.

[14] L. Fei-Fei, A. Iyer, C. Koch, P. Perona, What do we perceive in a glance of a real-world scene?, Journal of Vision 7 (1) (2007) 10–10.

[15] G. Rousselet, O. Joubert, M. Fabre-Thorpe, How long to get to the "gist" of real-world natural scenes?, Visual Cognition 12 (6) (2005) 852–877.

[16] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, International Journal of Computer Vision (IJCV) 42 (3) (2001) 145–175.

[17] A. Delorme, G. Richard, M. Fabre-Thorpe, Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans, Vision Research 40 (16) (2000) 2187–2200.

[18] T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization, Proceedings of the National Academy of Sciences (PNAS) 104 (15) (2007) 6424–6429.

[19] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), 2012.

[20] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, Y. L. Cun, Learning convolutional feature hierarchies for visual recognition, in: Advances in Neural Information Processing Systems (NeurIPS), 2010.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[22] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, J. J. DiCarlo, Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks, Journal of Neuroscience 38 (33) (2018) 7255–7269.

[23] A. Oliva, P. G. Schyns, Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli, Cognitive Psychology 34 (1) (1997) 72–107.

[24] P. G. Schyns, Diagnostic recognition: task constraints, object information, and their interactions, Cognition 67 (1-2) (1998) 147–179.

[25] G. L. Malcolm, A. Nuthmann, P. G. Schyns, Beyond gist: Strategic and incremental information accumulation for scene categorization, Psychological science 25 (5) (2014) 1087–1097.

[26] S. Qi, S. Huang, P. Wei, S.-C. Zhu, Predicting human activities using stochastic grammar, in: International Conference on Computer Vision (ICCV), 2017.

[27] M. Pei, Y. Jia, S.-C. Zhu, Parsing video events with goal inference and intent prediction, in: International Conference on Computer Vision (ICCV), 2011.

[28] F. Gosselin, P. G. Schyns, Bubbles: a technique to reveal the use of information in recognition tasks, Vision research 41 (17) (2001) 2261–2271.

[29] R. Hartley, A. Zisserman, Multiple view geometry in computer vision, Cambridge university press, 2003.

[30] Y. Ma, S. Soatto, J. Kosecka, S. S. Sastry, An invitation to 3-d vision: from images to geometric models, Vol. 26, Springer Science & Business Media, 2012.

[31] A. Gupta, M. Hebert, T. Kanade, D. M. Blei, Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces, in: Advances in Neural Information Processing Systems (NeurIPS), 2010.

[32] A. G. Schwing, S. Fidler, M. Pollefeys, R. Urtasun, Box in the box: Joint 3d layout and object reasoning from single images, in: International Conference on Computer Vision (ICCV), 2013.

[33] W. Choi, Y.-W. Chao, C. Pantofaru, S. Savarese, Understanding indoor scenes using 3d geometric phrases, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[34] Y. Zhao, S.-C. Zhu, Scene parsing by integrating function, geometry and appearance models, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[35] X. Liu, Y. Zhao, S.-C. Zhu, Single-view 3d scene reconstruction and parsing by attribute grammar, Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 40 (3) (2018) 710–725.

[36] S. Huang, S. Qi, Y. Zhu, Y. Xiao, Y. Xu, S.-C. Zhu, Holistic 3d scene parsing and reconstruction from a single rgb image, in: European Conference on Computer Vision (ECCV), 2018.

[37] E. C. Tolman, Cognitive maps in rats and men, Psychological review 55 (4) (1948) 189.

[38] R. F. Wang, E. S. Spelke, Comparative approaches to human navigation, The Neurobiology of Spatial Behaviour (2003) 119–143.

[39] J. J. Koenderink, A. J. van Doorn, A. M. Kappers, J. S. Lappin, Large-scale visual frontoparallels under full-cue conditions, Perception 31 (12) (2002) 1467–1475.

[40] W. H. Warren, D. B. Rothman, B. H. Schnapp, J. D. Ericson, Wormholes in virtual space: From cognitive maps to cognitive graphs, Cognition 166 (2017) 152–163.

[41] S. Gillner, H. A. Mallot, Navigation and acquisition of spatial knowledge in a virtual maze, Journal of Cognitive Neuroscience 10 (4) (1998) 445–463.

[42] P. Foo, W. H. Warren, A. Duchon, M. J. Tarr, Do humans integrate routes into a cognitive map? map-versus landmark-based navigation of novel shortcuts, Journal of Experimental Psychology: Learning, Memory, and Cognition 31 (2) (2005) 195.

[43] E. R. Chrastil, W. H. Warren, From cognitive maps to cognitive graphs, PloS one 9 (11) (2014) e112544.

[44] R. W. Byrne, Memory for urban geography, The Quarterly Journal of Experimental Psychology 31 (1) (1979) 147–154.

[45] B. Tversky, Distortions in cognitive maps, Geoforum 23 (2) (1992) 131–138.

[46] K. N. Ogle, Researches in binocular vision, WB Saunders, 1950.

[47] J. M. Foley, Binocular distance perception, Psychological review 87 (5) (1980) 411.

[48] R. K. Luneburg, Mathematical analysis of binocular vision, Princeton University Press, 1947.

[49] T. Indow, A critical review of luneburg's model with regard to global structure of visual space, Psychological review 98 (3) (1991) 430.

[50] W. C. Gogel, A theory of phenomenal geometry and its applications, Perception & Psychophysics 48 (2) (1990) 105–123.

[51] A. Glennerster, L. Tcheang, S. J. Gilson, A. W. Fitzgibbon, A. J. Parker, Humans ignore motion and stereo cues in favor of a fictional stable world, Current Biology 16 (4) (2006) 428–432.

[52] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, E. I. Moser, Microstructure of a spatial map in the entorhinal cortex, Nature 436 (7052) (2005) 801.

[53] N. J. Killian, M. J. Jutras, E. A. Buffalo, A map of visual space in the

primate entorhinal cortex, Nature 491 (7426) (2012) 761.
[54] J. O'keefe, L. Nadel, The hippocampus as a cognitive map, Oxford Clarendon Press, 1978.
[55] J. Jacobs, C. T. Weidemann, J. F. Miller, A. Solway, J. F. Burke, X.-X. Wei, N. Suthana, M. R. Sperling, A. D. Sharan, I. Fried, et al., Direct recordings of grid-like neuronal activity in human spatial navigation, Nature neuroscience 16 (9) (2013) 1188.
[56] M. Fyhn, T. Hafting, M. P. Witter, E. I. Moser, M.-B. Moser, Grid cells in mice, Hippocampus 18 (12) (2008) 1230–1238.
[57] C. F. Doeller, C. Barry, N. Burgess, Evidence for grid cells in a human memory network, Nature 463 (7281) (2010) 657.
[58] M. M. Yartsev, M. P. Witter, N. Ulanovsky, Grid cells without theta oscillations in the entorhinal cortex of bats, Nature 479 (7371) (2011) 103.
[59] R. Gao, J. Xie, S.-C. Zhu, Y. N. Wu, Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion, in: International Conference on Learning Representations (ICLR), 2019.
[60] L. Gootjes-Dreesbach, L. C. Pickup, A. W. Fitzgibbon, A. Glennerster, Comparison of view-based and reconstruction-based models of human navigational strategy, Journal of vision 17 (9) (2017) 11–11.
[61] J. Vuong, A. Fitzgibbon, A. Glennerster, Human pointing errors suggest a flattened, task-dependent representation of space, bioRxiv (2018) 390088.
[62] H. Choi, B. J. Scholl, Perceiving causality after the fact: Postdiction in the temporal dynamics of causal perception, Perception 35 (3) (2006) 385–399.
[63] B. J. Scholl, K. Nakayama, Illusory causal crescents: Misperceived spatial relations due to perceived causality, Perception 33 (4) (2004) 455–469.
[64] B. J. Scholl, T. Gao, Perceiving animacy and intentionality: Visual processing or higher-level judgment, Social perception: Detection and interpretation of animacy, agency, and intention 4629.
[65] B. J. Scholl, Objects and attention: The state of the art, Cognition 80 (1-2) (2001) 1–46.
[66] E. Vul, G. Alvarez, J. B. Tenenbaum, M. J. Black, Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model, in: Advances in Neural Information Processing Systems (NeurIPS), 2009.
[67] P. W. Battaglia, J. B. Hamrick, J. B. Tenenbaum, Simulation as an engine of physical scene understanding, Proceedings of the National Academy of Sciences (PNAS) 110 (45) (2013) 18327–18332.
[68] J. Hamrick, P. Battaglia, J. B. Tenenbaum, Internal physics models guide probabilistic judgments about object dynamics, in: Annual Meeting of the Cognitive Science Society (CogSci), 2011.
[69] D. Xie, T. Shu, S. Todorovic, S.-C. Zhu, Learning and inferring "dark matter" and predicting human intents and trajectories in videos, Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 40 (7) (2018) 1639–1652.
[70] T. Ullman, A. Stuhlmüller, N. Goodman, J. B. Tenenbaum, Learning physics from dynamical scenes, in: Annual Meeting of the Cognitive Science Society (CogSci), 2014.
[71] T. Gerstenberg, J. B. Tenenbaum, Intuitive theories, in: Oxford handbook of causal reasoning, Oxford University Press New York, NY, 2017, pp. 515–548.
[72] I. Newton, J. Colson, The Method of Fluxions and Infinite Series; with Its Application to the Geometry of Curve-lines, Henry Woodfall; and sold by John Nourse, 1736.
[73] C. Maclaurin, A Treatise of Fluxions: In Two Books. 1, Vol. 1, Ruddimans, 1742.
[74] E. T. Mueller, Commonsense reasoning: an event calculus based approach, Morgan Kaufmann, 2014.
[75] E. T. Mueller, Daydreaming in humans and machines: a computer model of the stream of thought, Intellect Books, 1990.
[76] A. Michotte, The perception of causality (TR Miles, Trans.), London, England: Methuen & Co, 1963.
[77] S. Carey, The origin of concepts, Oxford University Press, 2009.
[78] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
[79] D. Parikh, K. Grauman, Relative attributes, in: International Conference on Computer Vision (ICCV), 2011.
[80] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[81] B. Yao, S.-C. Zhu, Learning deformable action templates from cluttered videos, in: International Conference on Computer Vision (ICCV), 2009.
[82] B. Z. Yao, B. X. Nie, Z. Liu, S.-C. Zhu, Animated pose templates for modeling and detecting human actions, Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 36 (3) (2013) 436–452.
[83] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[84] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2005.
[85] S. Sadanand, J. J. Corso, Action bank: A high-level representation of activity in video, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[86] B. Zheng, Y. Zhao, C. Y. Joey, K. Ikeuchi, S.-C. Zhu, Detecting potential falling objects by inferring human action and natural disturbance, in: International Conference on Robotics and Automation (ICRA), 2014.
[87] R. Fleming, M. Barnett-Cowan, H. Bülthoff, Perceived object stability is affected by the internal representation of gravity, PLoS One 6 (4).
[88] M. Zago, F. Lacquaniti, Visual perception and interception of falling objects: a review of evidence for an internal model of gravity, Journal of Neural Engineering 2 (3) (2005) S198.
[89] P. J. Kellman, E. S. Spelke, Perception of partly occluded objects in infancy, Cognitive psychology 15 (4) (1983) 483–524.
[90] R. Baillargeon, E. S. Spelke, S. Wasserman, Object permanence in five-month-old infants, Cognition 20 (3) (1985) 191–208.
[91] S. P. Johnson, R. N. Aslin, Perception of object unity in 2-month-old infants, Developmental Psychology 31 (5) (1995) 739.
[92] A. Needham, Factors affecting infants' use of featural information in object segregation, Current Directions in Psychological Science 6 (2) (1997) 26–33.
[93] R. Baillargeon, Infants' physical world, Current directions in psychological science 13 (3) (2004) 89–94.
[94] B. Zheng, Y. Zhao, J. Yu, K. Ikeuchi, S.-C. Zhu, Scene understanding by reasoning stability and safety, International Journal of Computer Vision (IJCV) (2015) 221–238.
[95] S. Qi, Y. Zhu, S. Huang, C. Jiang, S.-C. Zhu, Human-centric indoor scene synthesis using stochastic grammar, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[96] S. Huang, S. Qi, Y. Xiao, Y. Zhu, Y. N. Wu, S.-C. Zhu, Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation, in: Advances in Neural Information Processing Systems (NeurIPS), 2018.
[97] M. Iacoboni, I. Molnar-Szakacs, V. Gallese, G. Buccino, J. C. Mazziotta, G. Rizzolatti, Grasping the intentions of others with one's own mirror neuron system, PLoS biology 3 (3) (2005) e79.
[98] A. Gupta, S. Satkin, A. A. Efros, M. Hebert, From 3d scene geometry to human workspace, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[99] G. Csibra, G. Gergely, 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans, Acta psychologica 124 (1) (2007) 60–78.
[100] C. L. Baker, J. B. Tenenbaum, R. R. Saxe, Goal inference as inverse planning, in: Annual Meeting of the Cognitive Science Society (CogSci), 2007.
[101] C. L. Baker, N. D. Goodman, J. B. Tenenbaum, Theory-based social goal inference, in: Annual Meeting of the Cognitive Science Society (CogSci), 2008.
[102] M. Hoai, F. De la Torre, Max-margin early event detectors, International Journal of Computer Vision (IJCV) 107 (2) (2014) 191–202.
[103] M. W. Turek, A. Hoogs, R. Collins, Unsupervised learning of functional categories in video scenes, in: European Conference on Computer Vision (ECCV), 2010.
[104] H. Grabner, J. Gall, L. Van Gool, What makes a chair a chair?, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[105] Z. Jia, A. Gallagher, A. Saxena, T. Chen, 3d-based reasoning with blocks, support, and stability, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[106] Y. Jiang, H. Koppula, A. Saxena, Hallucinated humans as the hidden context for labeling 3d scenes, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[107] T. Shu, S. M. Thurman, D. Chen, S.-C. Zhu, H. Lu, Critical features of joint actions that signal human interaction, in: Annual Meeting of the Cognitive Science Society (CogSci), 2016.
[108] T. Shu, Y. Peng, L. Fan, H. Lu, S.-C. Zhu, Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations, Topics in cognitive science 10 (1) (2018) 225–241.
[109] T. Shu, Y. Peng, H. Lu, S.-C. Zhu, Partitioning the perception of physical and social events within a unified psychological space, in: Annual Meeting of the Cognitive Science Society (CogSci), 2019.
[110] C. Baker, R. Saxe, J. Tenenbaum, Bayesian theory of mind: Modeling joint belief-desire attribution, in: Annual Meeting of the Cognitive Science Society (CogSci), 2011.
[111] Y. Zhao, S. Holtzen, T. Gao, S.-C. Zhu, Represent and infer human theory of mind for human-robot interaction, in: AAAI fall symposium series, 2015.
[112] A. A. Robb, Optical geometry of motion: A new view of the theory of relativity, W. Heffer, 1911.
[113] D. B. Malament, The class of continuous timelike curves determines the topology of spacetime, Journal of mathematical physics 18 (7) (1977) 1399–1404.
[114] A. A. Robb, Geometry of time and space, Cambridge University Press, 2014.
[115] R. Corrigan, P. Denton, Causal understanding as a developmental primitive, Developmental review 16 (2) (1996) 162–202.
[116] P. A. White, Causal processing: Origins and development, Psychological bulletin 104 (1) (1988) 36.
[117] Y.-C. Chen, B. J. Scholl, The perception of history: Seeing causal history in static shapes induces illusory motion perception, Psychological Science 27 (6) (2016) 923–930.
[118] K. Holyoak, P. W. Cheng, Causal learning and inference as a rational process: The new synthesis, Annual Review of Psychology 62 (2011) 135–163.
[119] D. R. Shanks, A. Dickinson, Associative accounts of causality judgment, Psychology of learning and motivation 21 (1988) 229–261.
[120] R. A. Rescorla, A. R. Wagner, A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement, Classical conditioning II: Current research and theory 2 (1972) 64–99.
[121] H. Lu, A. L. Yuille, M. Liljeholm, P. W. Cheng, K. J. Holyoak, Bayesian generic priors for causal learning, Psychological Review 115 (4) (2008) 955–984.
[122] M. Edmonds, S. Qi, Y. Zhu, J. Kubricht, S.-C. Zhu, H. Lu, Decomposing human causal learning: Bottom-up associative learning and top-down schema reasoning, in: Annual Meeting of the Cognitive Science Society (CogSci), 2019.
[123] M. R. Waldmann, K. J. Holyoak, Predictive and diagnostic learning within causal models: asymmetries in cue competition, Journal of Experimental Psychology: General 121 (2) (1992) 222–236.
[124] M. Edmonds, J. Kubricht, C. Summers, Y. Zhu, B. Rothrock, S.-C. Zhu, H. Lu, Human causal transfer: Challenges for deep reinforcement learning, in: Annual Meeting of the Cognitive Science Society (CogSci), 2018.
[125] P. W. Cheng, From covariation to causation: a causal power theory, Psychological Review 104 (2) (1997) 367–405.
[126] B. J. Scholl, P. D. Tremoulet, Perceptual causality and animacy, Trends in Cognitive Sciences 4 (8) (2000) 299–309.
[127] M. Rolfs, M. Dambacher, P. Cavanagh, Visual adaptation of the perception of causality, Current Biology 23 (3) (2013) 250–254.
[128] C. McCollough, Color adaptation of edge-detectors in the human visual system, Science 149 (3688) (1965) 1115–1116.
[129] J. Kominsky, B. Scholl, Retinotopically specific visual adaptation reveals the structure of causal events in perception, in: Annual Meeting of the Cognitive Science Society (CogSci), 2018.
[130] T. Gerstenberg, M. F. Peterson, N. D. Goodman, D. A. Lagnado, J. B. Tenenbaum, Eye-tracking causality, Psychological Science 28 (12) (2017) 1731–1744.
[131] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529.
[132] J. Schulman, S. Levine, P. Abbeel, M. Jordan, P. Moritz, Trust region policy optimization, in: International Conference on Machine Learning (ICML), 2015.
[133] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.
[134] S. Levine, C. Finn, T. Darrell, P. Abbeel, End-to-end training of deep visuomotor policies, The Journal of Machine Learning Research 17 (1) (2016) 1334–1373.
[135] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347.
[136] C. Zhang, O. Vinyals, R. Munos, S. Bengio, A study on overfitting in deep reinforcement learning, arXiv preprint arXiv:1804.06893.
[137] K. Kansky, T. Silver, D. A. Mély, M. Eldawy, M. Lázaro-Gredilla, X. Lou, N. Dorfman, S. Sidor, S. Phoenix, D. George, Schema networks: Zero-shot transfer with a generative causal model of intuitive physics, arXiv preprint arXiv:1706.04317.
[138] A. Fire, S.-C. Zhu, Learning perceptual causality from video, ACM Transactions on Intelligent Systems and Technology (TIST) 7 (2) (2016) 23.
[139] D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies., Journal of educational Psychology 66 (5) (1974) 688.
[140] G. W. Imbens, D. B. Rubin, Causal inference in statistics, social, and biomedical sciences, Cambridge University Press, 2015.
[141] P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.
[142] J. Pearl, Causality: Models, reasoning and inference, Cambridge University Press, 2000.
[143] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, T. Richardson, Causation, prediction, and search, MIT press, 2000.
[144] D. M. Chickering, Optimal structure identification with greedy search, Journal of machine learning research 3 (Nov) (2002) 507–554.
[145] J. Peters, J. M. Mooij, D. Janzing, B. Schölkopf, Causal discovery with continuous additive noise models, The Journal of Machine Learning Research 15 (1) (2014) 2009–2053.
[146] Y.-B. He, Z. Geng, Active learning of causal networks with intervention experiments and optimal designs, Journal of Machine Learning Research 9 (Nov) (2008) 2523–2547.
[147] N. R. Bramley, P. Dayan, T. L. Griffiths, D. A. Lagnado, Formalizing neurath's ship: Approximate algorithms for online causal learning, Psychological review 124 (3) (2017) 301.
[148] R. A. Fisher, The design of experiments, Oliver And Boyd; Edinburgh; London, 1937.
[149] A. Fire, S.-C. Zhu, Using causal induction in humans to learn and infer causality from video, in: Annual Meeting of the Cognitive Science Society (CogSci), 2013.
[150] S. C. Zhu, Y. N. Wu, D. Mumford, Minimax entropy principle and its application to texture modeling, Neural computation 9 (8) (1997) 1627–1660.
[151] Y. Xu, L. Qin, X. Liu, J. Xie, S.-C. Zhu, A causal and-or graph model for visibility fluent reasoning in tracking interacting objects, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[152] C. Xiong, N. Shukla, W. Xiong, S.-C. Zhu, Robot learning with a spatial, temporal, and causal and-or graph, in: International Conference on Robotics and Automation (ICRA), 2016.
[153] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, S.-C. Zhu, Beyond point clouds: Scene understanding by reasoning geometry and physics, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[154] M. McCloskey, A. Washburn, L. Felch, Intuitive physics: the straightdown belief and its origin, Journal of Experimental Psychology: Learning, Memory, and Cognition 9 (4) (1983) 636.
[155] M. McCloskey, A. Caramazza, B. Green, Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects, Science 210 (4474) (1980) 1139–1141.
[156] A. A. DiSessa, Unlearning aristotelian physics: A study of knowledge-based learning, Cognitive science 6 (1) (1982) 37–75.

[157] M. K. Kaiser, J. Jonides, J. Alexander, Intuitive reasoning about abstract and familiar physics problems, Memory & Cognition 14 (4) (1986) 308–312.

[158] K. A. Smith, P. Battaglia, E. Vul, Consistent physics underlying ballistic motion prediction, in: Annual Meeting of the Cognitive Science Society (CogSci), 2013.

[159] M. K. Kaiser, D. R. Proffitt, S. M. Whelan, H. Hecht, Influence of animation on dynamical judgments, Journal of experimental Psychology: Human Perception and performance 18 (3) (1992) 669.

[160] M. K. Kaiser, D. R. Proffitt, K. Anderson, Judgments of natural and anomalous trajectories in the presence and absence of motion, Journal of Experimental Psychology: Learning, Memory, and Cognition 11 (4) (1985) 795.

[161] I.-K. Kim, E. S. Spelke, Perception and understanding of effects of gravity and inertia on object motion, Developmental Science 2 (3) (1999) 339–362.

[162] P. Achinstein, The nature of explanation, Oxford University Press on Demand, 1983.

[163] J. Fischer, J. G. Mikhael, J. B. Tenenbaum, N. Kanwisher, Functional neuroanatomy of intuitive physical inference, Proceedings of the National Academy of Sciences (PNAS) 113 (34) (2016) E5072–E5081.

[164] Y. Chen, S. Huang, T. Yuan, Y. Zhu, S. Qi, S.-C. Zhu, Holistic++ scene understanding with human-object interaction and physical commonsense, in: International Conference on Computer Vision (ICCV), 2019.

[165] T. D. Ullman, E. Spelke, P. Battaglia, J. B. Tenenbaum, Mind games: Game engines as an architecture for intuitive physics, Trends in Cognitive Sciences 21 (9) (2017) 649–665.

[166] C. Bates, P. Battaglia, I. Yildirim, J. B. Tenenbaum, Humans predict liquid dynamics using probabilistic simulation, in: Annual Meeting of the Cognitive Science Society (CogSci), 2015.

[167] J. Kubricht, C. Jiang, Y. Zhu, S.-C. Zhu, D. Terzopoulos, H. Lu, Probabilistic simulation predicts human performance on viscous fluid-pouring problem, in: Annual Meeting of the Cognitive Science Society (CogSci), 2016.

[168] J. Kubricht, Y. Zhu, C. Jiang, D. Terzopoulos, S.-C. Zhu, H. Lu, Consistent probabilistic simulation underlying human judgment in substance dynamics, in: Annual Meeting of the Cognitive Science Society (CogSci), 2017.

[169] J. R. Kubricht, K. J. Holyoak, H. Lu, Intuitive physics: Current research and controversies, Trends in Cognitive Sciences 21 (10) (2017) 749–759.

[170] D. Mumford, A. Desolneux, Pattern theory: the stochastic analysis of real-world signals, AK Peters/CRC Press, 2010.

[171] D. Mumford, Pattern theory: a unifying perspective, in: First European congress of mathematics, Springer, 1994.

[172] B. Julesz, Visual pattern discrimination, IRE transactions on Information Theory 8 (2) (1962) 84–92.

[173] S. C. Zhu, Y. Wu, D. Mumford, Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling, International Journal of Computer Vision (IJCV) 27 (2) (1998) 107–126.

[174] B. Julesz, Textons, the elements of texture perception, and their interactions, Nature 290 (5802) (1981) 91.

[175] S.-C. Zhu, C.-E. Guo, Y. Wang, Z. Xu, What are textons?, International Journal of Computer Vision (IJCV) 62 (1-2) (2005) 121–143.

[176] C.-e. Guo, S.-C. Zhu, Y. N. Wu, Towards a mathematical theory of primal sketch and sketchability, in: International Conference on Computer Vision (ICCV), 2003.

[177] C.-e. Guo, S.-C. Zhu, Y. N. Wu, Primal sketch: Integrating structure and texture, Computer Vision and Image Understanding (CVIU) 106 (1) (2007) 5–19.

[178] M. Nitzberg, D. Mumford, The 2.1-d sketch, in: ICCV, 1990.

[179] J. Y. Wang, E. H. Adelson, Layered representation for motion analysis, in: Conference on Computer Vision and Pattern Recognition (CVPR), 1993.

[180] J. Y. Wang, E. H. Adelson, Representing moving images with layers, Transactions on Image Processing (TIP) 3 (5) (1994) 625–638.

[181] D. Marr, H. K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes, Proceedings of the Royal Society of London. Series B. Biological Sciences 200 (1140) (1978) 269–294.

[182] I. Binford, Visual perception by computer, in: IEEE Conference of Systems and Control, 1971.

[183] R. A. Brooks, Symbolic reasoning among 3-d models and 2-d images, Artificial Intelligence 17 (1-3) (1981) 285–348.

[184] T. Kanade, Recovery of the three-dimensional shape of an object from a single view, Artificial intelligence 17 (1-3) (1981) 409–460.

[185] D. Broadbent, A question of levels: Comment on McClelland and Rumelhart, American Psychological Association, 1985.

[186] D. Lowe, Perceptual organization and visual recognition, Vol. 5, Springer Science & Business Media, 2012.

[187] A. P. Pentland, Perceptual organization and the representation of natural form, in: Readings in Computer Vision, Elsevier, 1987, pp. 680–699.

[188] K. Koffka, Principles of Gestalt psychology, Routledge, 2013.

[189] D. Waltz, Understanding line drawings of scenes with shadows, in: The psychology of computer vision, 1975.

[190] H. G. Barrow, J. M. Tenenbaum, Interpreting line drawings as three-dimensional surfaces, Artificial Intelligence 17 (1-3) (1981) 75–116.

[191] D. G. Lowe, Three-dimensional object recognition from single two-dimensional images, Artificial Intelligence 31 (3) (1987) 355–395.

[192] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[193] R. L. Solso, M. K. MacLin, O. H. MacLin, Cognitive psychology, Pearson Education New Zealand, 2005.

[194] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, J. Tenenbaum, Galileo: Perceiving physical object properties by integrating a physics engine with deep learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2015.

[195] P. Dayan, G. E. Hinton, R. M. Neal, R. S. Zemel, The helmholtz machine, Neural computation 7 (5) (1995) 889–904.

[196] L. G. Roberts, Machine perception of three-dimensional solids, Ph.D. thesis, Massachusetts Institute of Technology (1963).

[197] I. Biederman, R. J. Mezzanotte, J. C. Rabinowitz, Scene perception: Detecting and judging objects undergoing relational violations, Cognitive psychology (1982) 143–177.

[198] M. Blum, A. Griffith, B. Neumann, A stability test for configurations of blocks, Tech. rep., Massachusetts Institute of Technology (1970).

[199] M. Brand, P. Cooper, L. Birnbaum, Seeing physics, or: Physics is for prediction, in: Proceedings of the Workshop on Physics-based Modeling in Computer Vision, 1995.

[200] A. Gupta, A. A. Efros, M. Hebert, Blocks world revisited: Image understanding using qualitative geometry and mechanics, in: European Conference on Computer Vision (ECCV), 2010.

[201] V. Hedau, D. Hoiem, D. Forsyth, Recovering the spatial layout of cluttered rooms, in: International Conference on Computer Vision (ICCV), 2009.

[202] D. C. Lee, M. Hebert, T. Kanade, Geometric reasoning for single image structure recovery, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[203] V. Hedau, D. Hoiem, D. Forsyth, Recovering free space of indoor scenes from a single image, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[204] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision (ECCV), 2012.

[205] A. G. Schwing, T. Hazan, M. Pollefeys, R. Urtasun, Efficient structured prediction for 3d indoor scene understanding, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[206] R. Guo, D. Hoiem, Support surface prediction in indoor scenes, in: International Conference on Computer Vision (ICCV), 2013.

[207] T. Shao, A. Monszpart, Y. Zheng, B. Koo, W. Xu, K. Zhou, N. J. Mitra, Imagining the unseen: Stability-based cuboid arrangements for scene understanding, ACM Transactions on Graphics (TOG) 33 (6).

[208] Y. Du, Z. Liu, H. Basevi, A. Leonardis, B. Freeman, J. Tenenbaum, J. Wu, Learning to exploit stability for 3d scene parsing, in: Advances in Neural Information Processing Systems (NeurIPS), 2018.

[209] Y. Zhu, Y. Zhao, S. Chun Zhu, Understanding tools: Task-oriented object modeling, learning and recognition, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[210] J. Wu, J. J. Lim, H. Zhang, J. B. Tenenbaum, W. T. Freeman, Physics 101: Learning physical object properties from unlabeled videos, in: British Machine Vision Conference (BMVC), 2016.

[211] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, S.-C. Zhu, Inferring forces and learning human utilities from videos, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[212] M. A. Brubaker, D. J. Fleet, The kneed walker for human pose tracking, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[213] M. A. Brubaker, L. Sigal, D. J. Fleet, Estimating contact dynamics, in: International Conference on Computer Vision (ICCV), 2009.

[214] M. A. Brubaker, D. J. Fleet, A. Hertzmann, Physics-based person tracking using the anthropomorphic walker, International Journal of Computer Vision (IJCV) 87 (1-2) (2010) 140.

[215] T.-H. Pham, A. Kheddar, A. Qammaz, A. A. Argyros, Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[216] Y. Wang, J. Min, J. Zhang, Y. Liu, F. Xu, Q. Dai, J. Chai, Video-based hand manipulation capture through composite motion control, ACM Transactions on Graphics (TOG) 32 (4) (2013) 43.

[217] W. Zhao, J. Zhang, J. Min, J. Chai, Robust realtime physics-based motion control for human grasping, ACM Transactions on Graphics (TOG) 32 (6) (2013) 207.

[218] J. J. Gibson, The perception of the visual world, Houghton Mifflin, 1950.

[219] J. J. Gibson, The senses considered as perceptual systems, Houghton Mifflin, 1966.

[220] K. Nelson, Concept, word, and sentence: interrelations in acquisition and development, Psychological review 81 (4) (1974) 267.

[221] J. J. Gibson, The theory of affordances, Hilldale, USA.

[222] M. Hassanin, S. Khan, M. Tahtali, Visual affordance and function understanding: A survey, arXiv preprint arXiv:1807.06775.

[223] H. Min, C. Yi, R. Luo, J. Zhu, S. Bi, Affordance research in developmental robotics: A survey, IEEE Transactions on Cognitive and Developmental Systems 8 (4) (2016) 237–255.

[224] J. Bohg, A. Morales, T. Asfour, D. Kragic, Data-driven grasp synthesis—a survey, IEEE Transactions on Robotics 30 (2) (2013) 289–309.

[225] N. Yamanobe, W. Wan, I. G. Ramirez-Alpizar, D. Petit, T. Tsuji, S. Akizuki, M. Hashimoto, K. Nagata, K. Harada, A brief review of affordance in robotic manipulation research, Advanced Robotics 31 (19-20) (2017) 1086–1101.

[226] W. Kohler, The mentality of apes, New York: Liverright, 1925.

[227] W. H. Thorpe, Learning and instinct in animals, Harvard University Press, 1956.

[228] K. P. Oakley, Man the tool-maker, University of Chicago Press, 1968.

[229] J. Goodall, The Chimpanzees of Gombe: Patterns of Behavior, Bellknap Press of the Harvard University Press, 1986.

[230] A. Whiten, J. Goodall, W. C. McGrew, T. Nishida, V. Reynolds, Y. Sugiyama, C. E. Tutin, R. W. Wrangham, C. Boesch, Cultures in chimpanzees, Nature 399 (6737) (1999) 682.

[231] R. W. Byrne, A. Whiten, Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans, Clarendon Press/Oxford University Press, 1988.

[232] G. Sabbatini, H. M. Manrique, C. Trapanese, A. D. B. Vizioli, J. Call, E. Visalberghi, Sequential use of rigid and pliable tools in tufted capuchin monkeys (sapajus spp.), Animal Behaviour 87 (2014) 213–220.

[233] G. R. Hunt, Manufacture and use of hook-tools by new caledonian crows, Nature 379 (6562) (1996) 249.

[234] A. A. Weir, J. Chappell, A. Kacelnik, Shaping of hooks in new caledonian crows, Science 297 (5583) (2002) 981–981.

[235] B. B. Beck, Animal tool behavior: The use and manufacture of tools by animals, Garland STPM Press New York, 1980.

[236] C. D. Bird, N. J. Emery, Insightful problem solving and creative tool modification by captive nontool-using rooks, Proceedings of the National Academy of Sciences (PNAS) 106 (25) (2009) 10370–10375.

[237] P. Freeman, A. Newell, A model for functional reasoning in design, in: International Joint Conference on Artificial Intelligence (IJCAI), 1971.

[238] P. H. Winston, Learning structural descriptions from examples, Tech. rep., Massachusetts Institute of Technology (1970).

[239] P. H. Winston, T. O. Binford, B. Katz, M. Lowry, Learning physical descriptions from functional definitions, examples, and precedents, in: AAAI Conference on Artificial Intelligence (AAAI), 1983.

[240] M. Brady, P. E. Agre, The mechanic's mate, in: Advances in Artificial Intelligence, Proceedings of the Sixth European Conference on Artificial Intelligence (ECAI), 1984.

[241] J. H. Connell, M. Brady, Generating and generalizing models of visual objects, Artificial Intelligence 31 (2) (1987) 159–183.

[242] S.-B. Ho, Representing and using functional definitions for visual recognition, Ph.D. thesis, The University of Wisconsin-Madison (1987).

[243] M. DiManzo, E. Trucco, F. Giunchiglia, F. Ricci, Fur: Understanding functional reasoning, International Journal of Intelligent Systems 4 (4) (1989) 431–457.

[244] M. Minsky, Society of mind, Simon and Schuster, 1988.

[245] L. Stark, K. Bowyer, Achieving generalized object recognition through reasoning about association of function to structure, Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 13 (10) (1991) 1097–1104.

[246] Z. Liu, W. T. Freeman, J. B. Tenenbaum, J. Wu, Physical primitive decomposition, in: European Conference on Computer Vision (ECCV), 2018.

[247] K. Fang, T.-L. Wu, D. Yang, S. Savarese, J. J. Lim, Demo2vec: Reasoning object affordances from online videos, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[248] C. Baber, Cognition and tool use: Forms of engagement in human and animal use of tools, CRC Press, 2003.

[249] B. Inhelder, J. Piaget, The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures, Vol. 22, Psychology Press, 1958.

[250] B. Strickland, B. J. Scholl, Visual perception involves event-type representations: The case of containment versus occlusion, Journal of Experimental Psychology: General 144 (3) (2015) 570.

[251] M. Casasola, L. B. Cohen, Infant categorization of containment, support and tight-fit spatial relationships, Developmental Science 5 (2) (2002) 247–264.

[252] S. J. Hespos, R. Baillargeon, Reasoning about containment events in very young infants, Cognition 78 (3) (2001) 207–245.

[253] S.-h. Wang, R. Baillargeon, S. Paterson, Detecting continuity violations in infancy: A new account and new evidence from covering and tube events, Cognition 95 (2) (2005) 129–173.

[254] S. J. Hespos, E. S. Spelke, Precursors to spatial language: The case of containment, in: The categorization of spatial entities in language and cognition, John Benjamins Publishing Company, 2007, pp. 233–245.

[255] E. Davis, G. Marcus, N. Frazier-Logue, Commonsense reasoning about containers using radically incomplete information, Artificial intelligence 248 (2017) 46–84.

[256] E. Davis, How does a box work? a study in the qualitative dynamics of solid objects, Artificial Intelligence 175 (1) (2011) 299–345.

[257] E. Davis, Pouring liquids: A study in commonsense physical reasoning, Artificial Intelligence 172 (12-13) (2008) 1540–1578.

[258] A. G. Cohn, Qualitative spatial representation and reasoning techniques, in: Annual Conference on Artificial Intelligence, Springer, 1997.

[259] A. G. Cohn, S. M. Hazarika, Qualitative spatial representation and reasoning: An overview, Fundamenta informaticae 46 (1-2) (2001) 1–29.

[260] W. Liang, Y. Zhao, Y. Zhu, S.-C. Zhu, Evaluating human cognition of containing relations with physical simulation, in: Annual Meeting of the Cognitive Science Society (CogSci), 2015.

[261] L.-F. Yu, N. Duncan, S.-K. Yeung, Fill and transfer: A simple physics-based approach for containability reasoning, in: International Conference on Computer Vision (ICCV), 2015.

[262] R. Mottaghi, C. Schenck, D. Fox, A. Farhadi, See the glass half full: Reasoning about liquid containers, their volume and content, in: International Conference on Computer Vision (ICCV), 2017.

[263] W. Liang, Y. Zhao, Y. Zhu, S.-C. Zhu, What is where: Inferring containment relations from videos, in: International Joint Conference on Artificial Intelligence (IJCAI), 2016.

[264] W. Liang, Y. Zhu, S.-C. Zhu, Tracking occluded objects and recovering incomplete trajectories by reasoning about containment relations and human actions, in: AAAI Conference on Artificial Intelligence (AAAI), 2018.

[265] Y. Jiang, M. Lim, A. Saxena, Learning object arrangements in 3d scenes using human context, in: International Conference on Machine Learning (ICML), 2012.

[266] C. Jiang, S. Qi, Y. Zhu, S. Huang, J. Lin, L.-F. Yu, D. Terzopoulos, S.-C. Zhu, Configurable 3d scene synthesis and 2d image rendering with per-pixel ground truth using stochastic grammars, International Journal

of Computer Vision (IJCV) (2018) 920–941.
[267] K. Dautenhahn, C. L. Nehaniv, Imitation in Animals and Artifacts, MIT Press Cambridge, MA, 2002.
[268] B. D. Argall, S. Chernova, M. Veloso, B. Browning, A survey of robot learning from demonstration, Robotics and Autonomous Systems 57 (5) (2009) 469–483.
[269] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters, et al., An algorithmic perspective on imitation learning, Foundations and Trends® in Robotics 7 (1-2) (2018) 1–179.
[270] Y. Gu, W. Sheng, M. Liu, Y. Ou, Fine manipulative action recognition through sensor fusion, in: International Conference on Intelligent Robots and Systems (IROS), 2015.
[271] F. L. Hammond, Y. Mengüç, R. J. Wood, Toward a modular soft sensor-embedded glove for human hand motion and tactile pressure measurement, in: International Conference on Intelligent Robots and Systems (IROS), 2014.
[272] H. Liu, X. Xie, M. Millar, M. Edmonds, F. Gao, Y. Zhu, V. J. Santos, B. Rothrock, S.-C. Zhu, A glove-based system for studying hand-object manipulation via joint pose and force sensing, in: International Conference on Intelligent Robots and Systems (IROS), 2017.
[273] M. Edmonds, F. Gao, X. Xie, H. Liu, S. Qi, Y. Zhu, B. Rothrock, S.-C. Zhu, Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles, in: International Conference on Intelligent Robots and Systems (IROS), 2017.
[274] H. Liu, Y. Zhang, W. Si, X. Xie, Y. Zhu, S.-C. Zhu, Interactive robot knowledge patching using augmented reality, in: International Conference on Robotics and Automation (ICRA), 2018.
[275] H. Liu, C. Zhang, Y. Zhu, C. Jiang, S.-C. Zhu, Mirroring without overimitation: Learning functionally equivalent manipulation actions, in: AAAI Conference on Artificial Intelligence (AAAI), 2019.
[276] D. C. Dennett, The intentional stance, MIT press, 1989.
[277] F. Heider, The psychology of interpersonal relations, Psychology Press, 2013.
[278] G. Gergely, Z. Nádasdy, G. Csibra, S. Bíró, Taking the intentional stance at 12 months of age, Cognition 56 (2) (1995) 165–193.
[279] D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind?, Behavioral and brain sciences 1 (4) (1978) 515–526.
[280] D. A. Baldwin, J. A. Baird, Discerning intentions in dynamic human action, Trends in Cognitive Sciences 5 (4) (2001) 171–178.
[281] A. L. Woodward, Infants selectively encode the goal object of an actor's reach, Cognition 69 (1) (1998) 1–34.
[282] A. N. Meltzoff, R. Brooks, Like me" as a building block for understanding other minds: Bodily acts, attention, and intention, Intentions and intentionality: Foundations of social cognition 171191.
[283] D. A. Baldwin, J. A. Baird, M. M. Saylor, M. A. Clark, Infants parse dynamic action, Child development 72 (3) (2001) 708–717.
[284] M. Tomasello, M. Carpenter, J. Call, T. Behne, H. Moll, Understanding and sharing intentions: The origins of cultural cognition, Behavioral and brain sciences 28 (5) (2005) 675–691.
[285] S. Biro, B. Hommel, Becoming an intentional agent: introduction to the special issue., Acta psychologica 124 (1) (2007) 1–7.
[286] G. Gergely, H. Bekkering, I. Király, Developmental psychology: Rational imitation in preverbal infants, Nature 415 (6873) (2002) 755.
[287] A. L. Woodward, J. A. Sommerville, S. Gerson, A. M. Henderson, J. Buresh, The emergence of intention attribution in infancy, Psychology of learning and motivation 51 (2009) 187–222.
[288] F. Heider, M. Simmel, An experimental study of apparent behavior, The American journal of psychology 57 (2) (1944) 243–259.
[289] M. Tomasello, Developing theories of intention (1999).
[290] P. Bloom, Intention, history, and artifact concepts, Cognition 60 (1) (1996) 1–29.
[291] T. Gao, G. E. Newman, B. J. Scholl, The psychophysics of chasing: A case study in the perception of animacy, Cognitive psychology 59 (2) (2009) 154–179.
[292] S. Holtzen, Y. Zhao, T. Gao, J. B. Tenenbaum, S.-C. Zhu, Inferring human intent from video by sampling hierarchical plans, in: International Conference on Intelligent Robots and Systems (IROS), 2016.
[293] D. S. Berry, S. J. Misovich, Methodological approaches to the study of social event perception, Personality and Social Psychology Bulletin 20 (2) (1994) 139–152.
[294] J. N. Bassili, Temporal and spatial contingencies in the perception of social events, Journal of Personality and Social Psychology 33 (6) (1976) 680.
[295] W. H. Dittrich, S. E. Lea, Visual perception of intentional motion, Perception 23 (3) (1994) 253–268.
[296] D. C. Dennett, Précis of the intentional stance, Behavioral and brain sciences 11 (3) (1988) 495–505.
[297] S. Liu, N. B. Brooks, E. S. Spelke, Origins of the concepts cause, cost, and goal in prereaching infants, Proceedings of the National Academy of Sciences (PNAS) (2019) 201904410.
[298] S. Liu, E. S. Spelke, Six-month-old infants expect agents to minimize the cost of their actions, Cognition 160 (2017) 35–42.
[299] G. Gergely, G. Csibra, Teleological reasoning in infancy: The naıve theory of rational action, Trends in Cognitive Sciences 7 (7) (2003) 287–292.
[300] C. L. Baker, R. Saxe, J. B. Tenenbaum, Action understanding as inverse planning, Cognition 113 (3) (2009) 329–349.
[301] L. M. Pereira, et al., Intention recognition via causal bayes networks plus plan generation, in: Portuguese Conference on Artificial Intelligence, Springer, 2009.
[302] S. Narang, A. Best, D. Manocha, Inferring user intent using bayesian theory of mind in shared avatar-agent virtual environments, IEEE Transactions on Visualization and Computer Graph (TVCG) 25 (5) (2019) 2113–2122.
[303] R. Nakahashi, C. L. Baker, J. B. Tenenbaum, Modeling human understanding of complex intentional action with a bayesian nonparametric subgoal model, in: AAAI Conference on Artificial Intelligence (AAAI), 2016.
[304] P. Wei, Y. Liu, T. Shu, N. Zheng, S.-C. Zhu, Where and why are they looking? jointly inferring human attention and intentions in complex tasks, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[305] Y. Kong, Y. Fu, Human action recognition and prediction: A survey, arXiv preprint arXiv:1806.11230.
[306] S.-J. Blakemore, J. Decety, From the perception of action to the understanding of intention, Nature reviews neuroscience 2 (8) (2001) 561.
[307] B. Elsner, B. Hommel, Effect anticipation and action control., Journal of experimental psychology: human perception and performance 27 (1) (2001) 229.
[308] B. Elsner, Infants' imitation of goal-directed actions: The role of movements and action effects, Acta psychologica 124 (1) (2007) 44–59.
[309] G. Rizzolatti, L. Craighero, The mirror-neuron system, Annual Review of Neuroscience 27 (2004) 169–192.
[310] J. T. Kaplan, M. Iacoboni, Getting a grip on other minds: Mirror neurons, intention understanding, and cognitive empathy, Social neuroscience 1 (3-4) (2006) 175–183.
[311] V. M. Reid, G. Csibra, J. Belsky, M. H. Johnson, Neural correlates of the perception of goal-directed action in infants, Acta psychologica 124 (1) (2007) 129–138.
[312] G. Csibra, G. Gergely, The teleological origins of mentalistic action explanations: A developmental hypothesis, Developmental Science 1 (2) (1998) 255–259.
[313] G. Gergely, The development of understanding self and agency, Blackwell handbook of childhood cognitive development (2002) 26–46.
[314] C. L. Kleinke, Gaze and eye contact: a research review, Psychological bulletin 100 (1) (1986) 78.
[315] N. J. Emery, The eyes have it: the neuroethology, function and evolution of social gaze, Neuroscience & Biobehavioral Reviews 24 (6) (2000) 581–604.
[316] J. K. Burgoon, L. K. Guerrero, K. Floyd, Nonverbal communication, Routledge, 2016.
[317] L. Fan, W. Wang, S. Huang, X. Tang, S.-C. Zhu, Understanding human gaze communication by spatio-temporal graph reasoning, in: International Conference on Computer Vision (ICCV), 2019.
[318] A. P. Melis, M. Tomasello, Chimpanzees (pan troglodytes) coordinate by communicating in a collaborative problem-solving task, Proceedings of the Royal Society B 286 (1901) (2019) 20190408.
[319] L. Fan, Y. Chen, P. Wei, W. Wang, S.-C. Zhu, Inferring shared attention in social scene videos, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[320] S. Trick, D. Koert, J. Peters, C. Rothkopf, Multimodal uncertainty reduction for intention recognition in human-robot interaction, arXiv preprint

arXiv:1907.02426.
[321] T. Shu, M. S. Ryoo, S.-C. Zhu, Learning social affordance for human-robot interaction, in: International Joint Conference on Artificial Intelligence (IJCAI), 2016.
[322] T. Shu, X. Gao, M. S. Ryoo, S.-C. Zhu, Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions, in: International Conference on Robotics and Automation (ICRA), 2017.
[323] C. Zhang, F. Gao, B. Jia, Y. Zhu, S.-C. Zhu, Raven: A dataset for relational and analogical visual reasoning, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[324] K. J. Holyoak, Analogy and relational reasoning, in: The Oxford Handbook of Thinking and Reasoning, Oxford University Press, 2012, pp. 234–259.
[325] J. C. e. a. Raven, Raven's progressive matrices, Western Psychological Services.
[326] M. Gao, X. Wang, K. Wu, A. Pradhana, E. Sifakis, C. Yuksel, C. Jiang, Gpu optimization of material point methods, ACM Transactions on Graphics (TOG) 37 (6).
[327] N. Nisan, A. Ronen, Algorithmic mechanism design, Games and Economic behavior 35 (1-2) (2001) 166–196.
[328] J. Bentham, An introduction to the principles of morals, London: Athlone.
[329] K. J. Holyoak, P. Thagard, The analogical mind, American psychologist 52 (1) (1997) 35.
[330] P. W. Cheng, M. J. Buehner, Causal learning, in: The Oxford Handbook of Thinking and Reasoning, Oxford University Press, 2012, pp. 210–233.
[331] M. B. Hesse, Models and analogies in science, Notre Dame University Press, 1966.
[332] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems (NeurIPS), 2013.
[333] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.
[334] P. A. Carpenter, M. A. Just, P. Shell, What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test, Psychological review 97 (3) (1990) 404.
[335] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, Vqa: Visual question answering, in: International Conference on Computer Vision (ICCV), 2015.
[336] R. E Snow, P. Kyllonen, B. Marshalek, The topography of ability and learning correlations, Advances in the psychology of human intelligence (1984) 47–103.
[337] S. M. Jaeggi, M. Buschkuehl, J. Jonides, W. J. Perrig, Improving fluid intelligence with training on working memory, Proceedings of the National Academy of Sciences (PNAS) 105 (19) (2008) 6829–6833.
[338] G. H. Bower, A contrast effect in differential conditioning, Journal of Experimental Psychology 62 (2) (1961) 196.
[339] D. R. Meyer, The effects of differential rewards on discrimination reversal learning by monkeys, Journal of Experimental Psychology 41 (4) (1951) 268.
[340] A. M. Schrier, H. F. Harlow, Effect of amount of incentive on discrimination learning by monkeys, Journal of comparative and physiological psychology 49 (2) (1956) 117.
[341] R. M. Shapley, J. D. Victor, The effect of contrast on the transfer properties of cat retinal ganglion cells, The Journal of physiology 285 (1) (1978) 275–298.
[342] R. Lawson, Brightness discrimination performance and secondary reward strength as a function of primary reward amount, Journal of Comparative and Physiological Psychology 50 (1) (1957) 35.
[343] A. Amsel, Frustrative nonreward in partial reinforcement and discrimination learning: Some recent history and a theoretical extension, Psychological review 69 (4) (1962) 306.
[344] J. J. Gibson, E. J. Gibson, Perceptual learning: Differentiation or enrichment?, Psychological review 62 (1) (1955) 32.
[345] J. J. Gibson, The ecological approach to visual perception: classic edition, Psychology Press, 2014.
[346] R. Catrambone, K. J. Holyoak, Overcoming contextual limitations on problem-solving transfer, Journal of Experimental Psychology: Learning, Memory, and Cognition 15 (6) (1989) 1147.
[347] D. Gentner, V. Gunn, Structural alignment facilitates the noticing of differences, Memory & Cognition 29 (4) (2001) 565–577.
[348] R. Hammer, G. Diesendruck, D. Weinshall, S. Hochstein, The development of category learning strategies: What makes the difference?, Cognition 112 (1) (2009) 105–119.
[349] M. L. Gick, K. Paterson, Do contrasting examples facilitate schema acquisition and analogical transfer?, Canadian Journal of Psychology/Revue canadienne de psychologie 46 (4) (1992) 539.
[350] E. Haryu, M. Imai, H. Okada, Object similarity bootstraps young children to action-based verb extension, Child Development 82 (2) (2011) 674–686.
[351] L. Smith, D. Gentner, The role of difference-detection in learning contrastive categories, in: Annual Meeting of the Cognitive Science Society (CogSci), 2014.
[352] D. Gentner, Structure-mapping: A theoretical framework for analogy, Cognitive science 7 (2) (1983) 155–170.
[353] D. Gentner, A. B. Markman, Structural alignment in comparison: No difference without similarity, Psychological science 5 (3) (1994) 152–158.
[354] D. L. Schwartz, C. C. Chase, M. A. Oppezzo, D. B. Chin, Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer, Journal of Educational Psychology 103 (4) (2011) 759.
[355] C. Zhang, B. Jia, F. Gao, Y. Zhu, H. Lu, S.-C. Zhu, Learning perceptual inference by contrasting, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.
[356] D. Terzopoulos, J. Platt, A. Barr, K. Fleischer, Elastically deformable models, ACM Transactions on Graphics (TOG) 21 (4) (1987) 205–214.
[357] D. Terzopoulos, K. Fleischer, Modeling inelastic deformation: viscoelasticity, plasticity, fracture, ACM Transactions on Graphics (TOG) 22 (4) (1988) 269–278.
[358] N. Foster, D. Metaxas, Realistic animation of liquids, Graphical models and image processing 58 (5) (1996) 471–483.
[359] J. Stam, Stable fluids, in: ACM Transactions on Graphics (TOG), Vol. 99, 1999.
[360] R. Bridson, Fluid simulation for computer graphics, CRC Press, 2015.
[361] J. Bonet, R. D. Wood, Nonlinear continuum mechanics for finite element analysis, Cambridge university press, 1997.
[362] S. Blemker, J. Teran, E. Sifakis, R. Fedkiw, S. Delp, Fast 3d muscle simulations using a new quasistatic invertible finite-element algorithm, in: International Symposium on Computer Simulation in Biomechanics, 2005.
[363] J. Hegemann, C. Jiang, C. Schroeder, J. M. Teran, A level set method for ductile fracture, in: ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA), 2013.
[364] T. F. Gast, C. Schroeder, A. Stomakhin, C. Jiang, J. M. Teran, Optimization integrator for large time steps, IEEE Transactions on Visualization and Computer Graph (TVCG) 21 (10) (2015) 1103–1115.
[365] M. Li, M. Gao, T. Langlois, C. Jiang, D. M. Kaufman, Decomposed optimization time integrator for large-step elastodynamics, ACM Transactions on Graphics (TOG) 38 (4) (2019) 70.
[366] Y. Wang, C. Jiang, C. Schroeder, J. Teran, An adaptive virtual node algorithm with robust mesh cutting, in: ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA), 2014.
[367] J. J. Monaghan, Smoothed particle hydrodynamics, Annual review of astronomy and astrophysics 30 (1) (1992) 543–574.
[368] W. K. Liu, S. Jun, Y. F. Zhang, Reproducing kernel particle methods, International journal for numerical methods in fluids 20 (8-9) (1995) 1081–1106.
[369] S. Li, W. K. Liu, Meshfree and particle methods and their applications, Applied Mechanics Reviews 55 (1) (2002) 1–34.
[370] J. Donea, S. Giuliani, J.-P. Halleux, An arbitrary lagrangian-eulerian finite element method for transient dynamic fluid-structure interactions, Computer methods in applied mechanics and engineering 33 (1-3) (1982) 689–723.
[371] X. Xie, H. Liu, Z. Zhang, Y. Qiu, F. Gao, S. Qi, Y. Zhu, S.-C. Zhu, Vrgym: A virtual testbed for physical and interactive ai, in: Proceedings of the ACM TURC, 2019.
[372] J. U. Brackbill, H. M. Ruppel, Flip: A method for adaptively zoned, particle-in-cell calculations of fluid flows in two dimensions, Journal of Computational physics 65 (2) (1986) 314–343.

[373] C. Jiang, C. Schroeder, A. Selle, J. Teran, A. Stomakhin, The affine particle-in-cell method, ACM Transactions on Graphics (TOG) 34 (4) (2015) 51.

[374] D. Sulsky, Z. Chen, H. L. Schreyer, A particle method for history-dependent materials, Computer methods in applied mechanics and engineering 118 (1-2) (1994) 179–196.

[375] D. Sulsky, S.-J. Zhou, H. L. Schreyer, Application of a particle-in-cell method to solid mechanics, Computer physics communications 87 (1-2) (1995) 236–252.

[376] A. Stomakhin, C. Schroeder, L. Chai, J. Teran, A. Selle, A material point method for snow simulation, ACM Transactions on Graphics (TOG) 32 (4) (2013) 102.

[377] J. Gaume, T. Gast, J. Teran, A. van Herwijnen, C. Jiang, Dynamic anti-crack propagation in snow, Nature communications 9 (1) (2018) 3047.

[378] D. Ram, T. Gast, C. Jiang, C. Schroeder, A. Stomakhin, J. Teran, P. Kavehpour, A material point method for viscoelastic fluids, foams and sponges, in: ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA), 2015.

[379] Y. Yue, B. Smith, C. Batty, C. Zheng, E. Grinspun, Continuum foam: A material point method for shear-dependent flows, ACM Transactions on Graphics (TOG) 34 (5) (2015) 160.

[380] Y. Fang, M. Li, M. Gao, C. Jiang, Silly rubber: an implicit material point method for simulating non-equilibrated viscoelastic and elasto-plastic solids, ACM Transactions on Graphics (TOG) 38 (4) (2019) 118.

[381] G. Klar, T. Gast, A. Pradhana, C. Fu, C. Schroeder, C. Jiang, J. Teran, Drucker-prager elastoplasticity for sand animation, ACM Transactions on Graphics (TOG) 35 (4) (2016) 103.

[382] G. Daviet, F. Bertails-Descoubes, A semi-implicit material point method for the continuum simulation of granular materials, ACM Transactions on Graphics (TOG) 35 (4) (2016) 102.

[383] Y. Hu, Y. Fang, Z. Ge, Z. Qu, Y. Zhu, A. Pradhana, C. Jiang, A moving least squares material point method with displacement discontinuity and two-way rigid body coupling, ACM Transactions on Graphics (TOG) 37 (4) (2018) 150.

[384] S. Wang, M. Ding, T. F. Gast, L. Zhu, S. Gagniere, C. Jiang, J. M. Teran, Simulation and visualization of ductile fracture with the material point method, ACM Transactions on Graphics (TOG) 2 (2) (2019) 18.

[385] J. Wolper, Y. Fang, M. Li, J. Lu, M. Gao, C. Jiang, Cd-mpm: continuum damage material point methods for dynamic fracture animation, ACM Transactions on Graphics (TOG) 38 (4) (2019) 119.

[386] C. Jiang, T. Gast, J. Teran, Anisotropic elastoplasticity for cloth, knit and hair frictional contact, ACM Transactions on Graphics (TOG) 36 (4) (2017) 152.

[387] X. Han, T. F. Gast, Q. Guo, S. Wang, C. Jiang, J. Teran, A hybrid material point method for frictional contact with diverse materials, ACM Transactions on Graphics (TOG) 2 (2) (2019) 17.

[388] C. Fu, Q. Guo, T. Gast, C. Jiang, J. Teran, A polynomial particle-in-cell method, ACM Transactions on Graphics (TOG) 36 (6) (2017) 222.

[389] A. Stomakhin, C. Schroeder, C. Jiang, L. Chai, J. Teran, A. Selle, Augmented mpm for phase-change and varied materials, ACM Transactions on Graphics (TOG) 33 (4) (2014) 138.

[390] A. P. Tampubolon, T. Gast, G. Klár, C. Fu, J. Teran, C. Jiang, K. Museth, Multi-species simulation of porous sand and water mixtures, ACM Transactions on Graphics (TOG) 36 (4) (2017) 105.

[391] M. Gao, A. Pradhana, X. Han, Q. Guo, G. Kot, E. Sifakis, C. Jiang, Animating fluid sediment mixture in particle-laden flows, ACM Transactions on Graphics (TOG) 37 (4) (2018) 149.

[392] J. A. Nairn, Material point method calculations with explicit cracks, Computer Modeling in Engineering and Sciences 4 (6) (2003) 649–664.

[393] Z. Chen, L. Shen, Y.-W. Mai, Y.-G. Shen, A bifurcation-based decohesion model for simulating the transition from localization to decohesion with the mpm, Zeitschrift für Angewandte Mathematik und Physik (ZAMP) 56 (5) (2005) 908–930.

[394] H. Schreyer, D. Sulsky, S.-J. Zhou, Modeling delamination as a strong discontinuity with the material point method, Computer Methods in Applied Mechanics and Engineering 191 (23) (2002) 2483–2507.

[395] D. Sulsky, H. L. Schreyer, Axisymmetric form of the material point method with applications to upsetting and taylor impact problems, Computer Methods in Applied Mechanics and Engineering 139 (1-4) (1996) 409–429.

[396] P. Huang, X. Zhang, S. Ma, H. Wang, Shared memory openmp parallelization of explicit mpm and its application to hypervelocity impact, CMES: Computer Modelling in Engineering & Sciences 38 (2) (2008) 119–148.

[397] W. Hu, Z. Chen, Model-based simulation of the synergistic effects of blast and fragmentation on a concrete wall using the mpm, International journal of impact engineering 32 (12) (2006) 2066–2096.

[398] A. R. York, D. Sulsky, H. L. Schreyer, Fluid–membrane interaction based on the material point method, International Journal for Numerical Methods in Engineering 48 (6) (2000) 901–924.

[399] S. Bandara, K. Soga, Coupling of soil deformation and pore fluid flow using material point method, Computers and geotechnics 63 (2015) 199–214.

[400] J. E. Guilkey, J. B. Hoying, J. A. Weiss, Computational modeling of multicellular constructs with the material point method, Journal of biomechanics 39 (11) (2006) 2074–2086.

[401] P. HUANG, Material point method for metal and soil impact dynamics problems, Tsinghua University, 2010.

[402] Y. Fang, Y. Hu, S.-M. Hu, C. Jiang, A temporally adaptive material point method with regional time stepping, in: Computer Graphics Forum, 2018.

[403] S. Bardenhagen, E. Kober, The generalized interpolation material point method, Computer Modeling in Engineering and Sciences 5 (6) (2004) 477–496.

[404] M. Gao, A. P. Tampubolon, C. Jiang, E. Sifakis, An adaptive generalized interpolation material point method for simulating elastoplastic materials, ACM Transactions on Graphics (TOG) 36 (6) (2017) 223.

[405] A. Sadeghirad, R. M. Brannon, J. Burghardt, A convected particle domain interpolation technique to extend applicability of the material point method for problems involving massive deformations, International Journal for numerical methods in Engineering 86 (12) (2011) 1435–1456.

[406] D. Z. Zhang, X. Ma, P. T. Giguere, Material point method enhanced by modified gradient of shape function, Journal of Computational Physics 230 (16) (2011) 6379–6398.

[407] S. Legg, M. Hutter, Universal intelligence: A definition of machine intelligence, Minds and machines 17 (4) (2007) 391–444.

[408] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, H. Su, Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[409] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al., Shapenet: An information-rich 3d model repository, arXiv preprint arXiv:1512.03012.

[410] T. Feng, L.-F. Yu, S.-K. Yeung, K. Yin, K. Zhou, Crowd-driven mid-scale layout design., ACM Transactions on Graphics (TOG) 35 (4) (2016) 132–1.

[411] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, V. Koltun, Minos: Multimodal indoor simulator for navigation in complex environments, arXiv preprint arXiv:1712.03931.

[412] S. Brodeur, E. Perez, A. Anand, F. Golemo, L. Celotti, F. Strub, J. Rouat, H. Larochelle, A. Courville, Home: A household multimodal environment, arXiv preprint arXiv:1711.11017.

[413] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, S. Savarese, Gibson env: Real-world perception for embodied agents, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[414] Y. Wu, Y. Wu, G. Gkioxari, Y. Tian, Building generalizable agents with a realistic and rich 3d environment, arXiv preprint arXiv:1801.02209.

[415] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, A. Farhadi, Ai2-thor: An interactive 3d environment for visual ai, arXiv preprint arXiv:1712.05474.

[416] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, A. Torralba, Virtualhome: Simulating household activities via programs, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[417] X. Gao, R. Gong, T. Shu, X. Xie, S. Wang, S.-C. Zhu, Vrkitchen: an interactive 3d virtual environment for task-oriented learning, arXiv preprint arXiv:1903.05757.

[418] S. Shah, D. Dey, C. Lovett, A. Kapoor, Airsim: High-fidelity visual and physical simulation for autonomous vehicles, in: Field and service robotics, Springer, 2018.

[419] N. Shukla, Utility learning, non-markovian planning, and task-oriented programming language, Ph.D. thesis, UCLA (2019).

[420] N. Shukla, Y. He, F. Chen, S.-C. Zhu, Learning human utility from video demonstrations for deductive planning in robotics, in: Conference on Robot Learning, 2017.
[421] H. P. Grice, P. Cole, J. Morgan, et al., Logic and conversation, 1975 (1975) 41–58.
[422] N. D. Goodman, M. C. Frank, Pragmatic language interpretation as probabilistic inference, Trends in Cognitive Sciences 20 (11) (2016) 818–829.
[423] M. C. Frank, N. D. Goodman, Predicting pragmatic reasoning in language games, Science 336 (6084) (2012) 998–998.
[424] D. Lewis, Convention: A philosophical study, John Wiley & Sons, 2008.
[425] D. Sperber, D. Wilson, Relevance: Communication and cognition, Vol. 142, Harvard University Press Cambridge, MA, 1986.
[426] L. Wittgenstein, Philosophical investigations.
[427] H. H. Clark, Using language, Cambridge university press, 1996.
[428] C. Qing, M. Franke, Variations on a bayesian theme: Comparing bayesian models of referential reasoning, in: Bayesian natural language semantics and pragmatics, Springer, 2015, pp. 201–220.
[429] N. D. Goodman, A. Stuhlmüller, Knowledge and implicature: Modeling language understanding as social cognition, Topics in cognitive science 5 (1) (2013) 173–184.
[430] R. Dale, E. Reiter, Computational interpretations of the gricean maxims in the generation of referring expressions, Cognitive science 19 (2) (1995) 233–263.
[431] A. Benz, G. Jäger, R. Van Rooij, An introduction to game theory for linguists, in: Game theory and pragmatics, Springer, 2006, pp. 1–82.
[432] G. Jäger, Applications of game theory in linguistics, Language and Linguistics compass 2 (3) (2008) 406–421.
[433] M. Kinney, C. Tsatsoulis, Learning communication strategies in multiagent systems, Applied intelligence 9 (1) (1998) 71–91.
[434] D. S. Bernstein, R. Givan, N. Immerman, S. Zilberstein, The complexity of decentralized control of markov decision processes, Mathematics of operations research 27 (4) (2002) 819–840.
[435] C. V. Goldman, S. Zilberstein, Optimizing information exchange in cooperative multi-agent systems, in: International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2003.
[436] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing atari with deep reinforcement learning, arXiv preprint arXiv:1312.5602.
[437] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, R. Vicente, Multiagent cooperation and competition with deep reinforcement learning, PloS one 12 (4) (2017) e0172395.
[438] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. Torr, P. Kohli, S. Whiteson, Stabilising experience replay for deep multi-agent reinforcement learning, in: International Conference on Machine Learning (ICML), 2017.
[439] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, in: Advances in Neural Information Processing Systems (NeurIPS), 2017.
[440] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, S. Whiteson, Counterfactual multi-agent policy gradients, in: AAAI Conference on Artificial Intelligence (AAAI), 2018.
[441] K. Atkinson, T. Bench-Capon, P. McBurney, A dialogue game protocol for multi-agent argument over proposals for action, Autonomous Agents and Multi-Agent Systems 11 (2) (2005) 153–171.
[442] M. D. Sadek, P. Bretier, F. Panaget, Artimis: Natural dialogue meets rational agency, in: International Joint Conference on Artificial Intelligence (IJCAI), 1997.
[443] Z. Zheng, W. Wang, S. Qi, S.-C. Zhu, Reasoning visual dialogs with structural and partial observations, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[444] J. Foerster, I. A. Assael, N. de Freitas, S. Whiteson, Learning to communicate with deep multi-agent reinforcement learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2016.
[445] S. Sukhbaatar, R. Fergus, et al., Learning multiagent communication with backpropagation, in: Advances in Neural Information Processing Systems (NeurIPS), 2016.
[446] I. Mordatch, P. Abbeel, Emergence of grounded compositional language in multi-agent populations, in: AAAI Conference on Artificial Intelligence (AAAI), 2018.
[447] A. Lazaridou, A. Peysakhovich, M. Baroni, Multi-agent cooperation and the emergence of (natural) language, in: International Conference on Learning Representations (ICLR), 2017.
[448] S. Havrylov, I. Titov, Emergence of language with multi-agent games: Learning to communicate with sequences of symbols, in: Advances in Neural Information Processing Systems (NeurIPS), 2017.
[449] K. Evtimova, A. Drozdov, D. Kiela, K. Cho, Emergent language in a multi-modal, multi-step referential game, arXiv preprint arXiv:1705.10369.
[450] A. Lazaridou, K. M. Hermann, K. Tuyls, S. Clark, Emergence of linguistic communication from referential games with symbolic and pixel input, in: International Conference on Learning Representations (ICLR), 2018.
[451] K. Wagner, J. A. Reggia, J. Uriagereka, G. S. Wilkinson, Progress in the simulation of emergent communication and language, Adaptive Behavior 11 (1) (2003) 37–69.
[452] R. Ibsen-Jensen, J. Tkadlec, K. Chatterjee, M. A. Nowak, Language acquisition with communication between learners, Journal of The Royal Society Interface 15 (140) (2018) 20180073.
[453] L. Graesser, K. Cho, D. Kiela, Emergent linguistic phenomena in multi-agent communication games, arXiv preprint arXiv:1901.08706.
[454] E. Dupoux, P. Jacob, Universal moral grammar: a critical appraisal, Trends in Cognitive Sciences 11 (9) (2007) 373–378.
[455] J. Mikhail, Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment, Cambridge University Press, 2011.
[456] M. Kleiman-Weiner, R. Saxe, J. B. Tenenbaum, Learning a commonsense moral theory, cognition 167 (2017) 107–123.
[457] P. Blake, K. McAuliffe, J. Corbit, T. Callaghan, O. Barry, A. Bowie, L. Kleutsch, K. Kramer, E. Ross, H. Vongsachang, et al., The ontogeny of fairness in seven societies, Nature 528 (7581) (2015) 258.
[458] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, In search of homo economicus: behavioral experiments in 15 small-scale societies, American Economic Review 91 (2) (2001) 73–78.
[459] B. R. House, J. B. Silk, J. Henrich, H. C. Barrett, B. A. Scelza, A. H. Boyette, B. S. Hewlett, R. McElreath, S. Laurence, Ontogeny of prosocial behavior across diverse societies, Proceedings of the National Academy of Sciences (PNAS) 110 (36) (2013) 14586–14591.
[460] J. Graham, P. Meindl, E. Beall, K. M. Johnson, L. Zhang, Cultural differences in moral judgment and behavior, across and within societies, Current Opinion in Psychology 8 (2016) 125–130.
[461] T. Hurka, Virtue, vice, and value, Oxford University Press, 2000.
[462] J. Rawls, A theory of justice, Harvard university press, 1971.
[463] J. Haidt, The new synthesis in moral psychology, science 316 (5827) (2007) 998–1002.
[464] J. K. Hamlin, Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core, Current Directions in Psychological Science 22 (3) (2013) 186–193.
[465] R. Kim, M. Kleiman-Weiner, A. Abeliuk, E. Awad, S. Dsouza, J. B. Tenenbaum, I. Rahwan, A computational model of commonsense moral decision making, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2018.
[466] M. Kleiman-Weiner, T. Gerstenberg, S. Levine, J. B. Tenenbaum, Inference of intention and permissibility in moral decision making., in: Annual Meeting of the Cognitive Science Society (CogSci), 2015.