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]*Center for Vision, Cognition, Learning, and Autonomy (VCLA), UCLA*
[b]*Center for Brains, Minds, and Machines (CBMM), MIT*

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, causality, and utility (FPICU), as the five core domains of the cognitive AI with human-like commonsense. FPICU 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 FPICU as dark matters. We demonstrate the power of this cognitive AI approach with human-like commonsense by showing how to apply FPICU 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" human-like commonsense for solving novel tasks.

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

1. Call for a Paradigm Shift in Vision and AI

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 labeled 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 matter 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 awareness, in the vision and broader communities, 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;

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

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

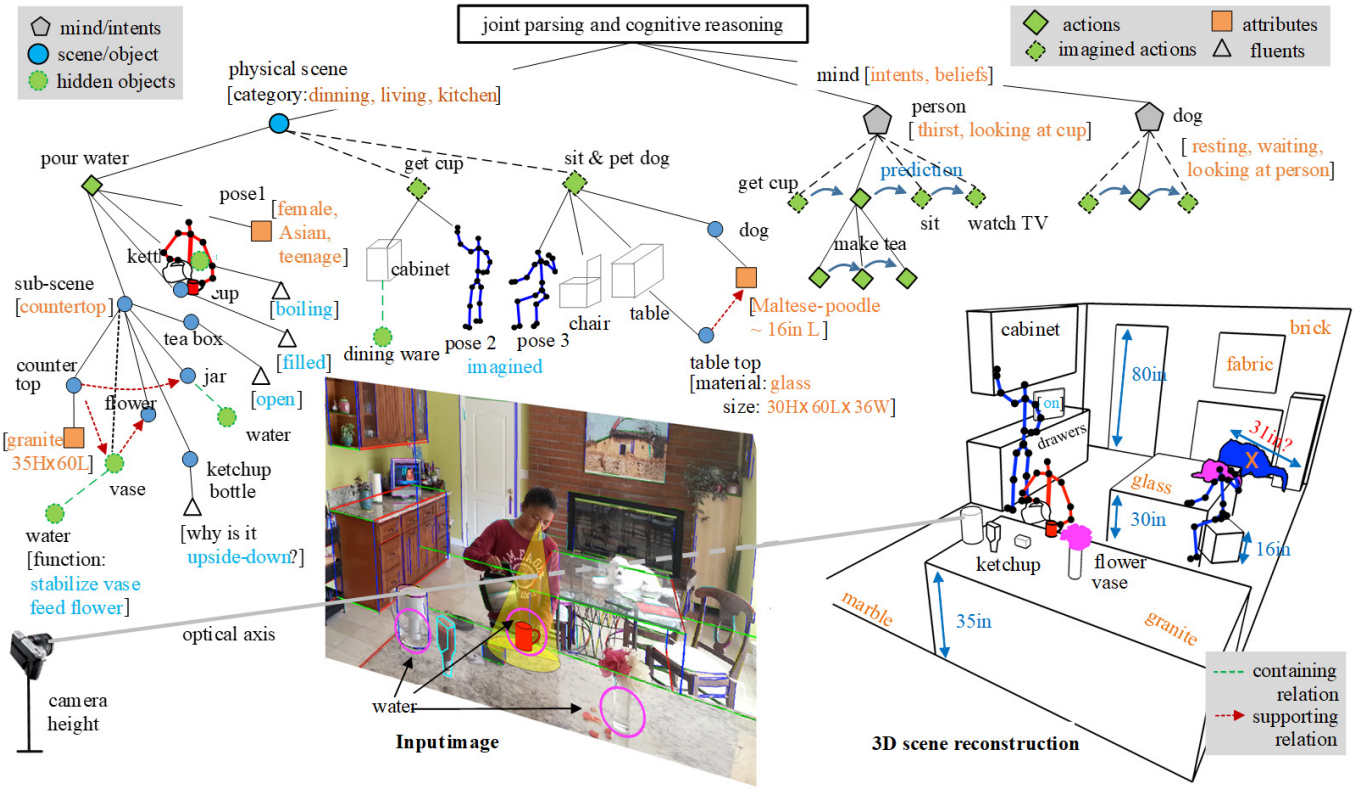


Figure 1: An example of an in-depth understanding of the scene and event by joint parsing and cognitive reasoning. From a single image, a computer vision system should be able to jointly (1) reconstruct the 3D scene, (2) estimate camera parameters, materials, and illumination, (3) parse the scene hierarchically 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 AI with human-like core knowledge, which are largely missing in the current computer vision research.

see an example 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, intents/goals) 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 the 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. Take Figure 1 as an example; the gender of the agent is "hidden". In contrast, the functionality of a kettle is "dark"; one can infer that there is liquid inside it. The pose of the ketchup bottle could also be considered as "dark" as we intentionally put the bottle upside down to utilize gravity for easy use.

In the remainder of the paper, we start by revisiting a classic view of computer vision in terms of "what" is "where" in Sec-

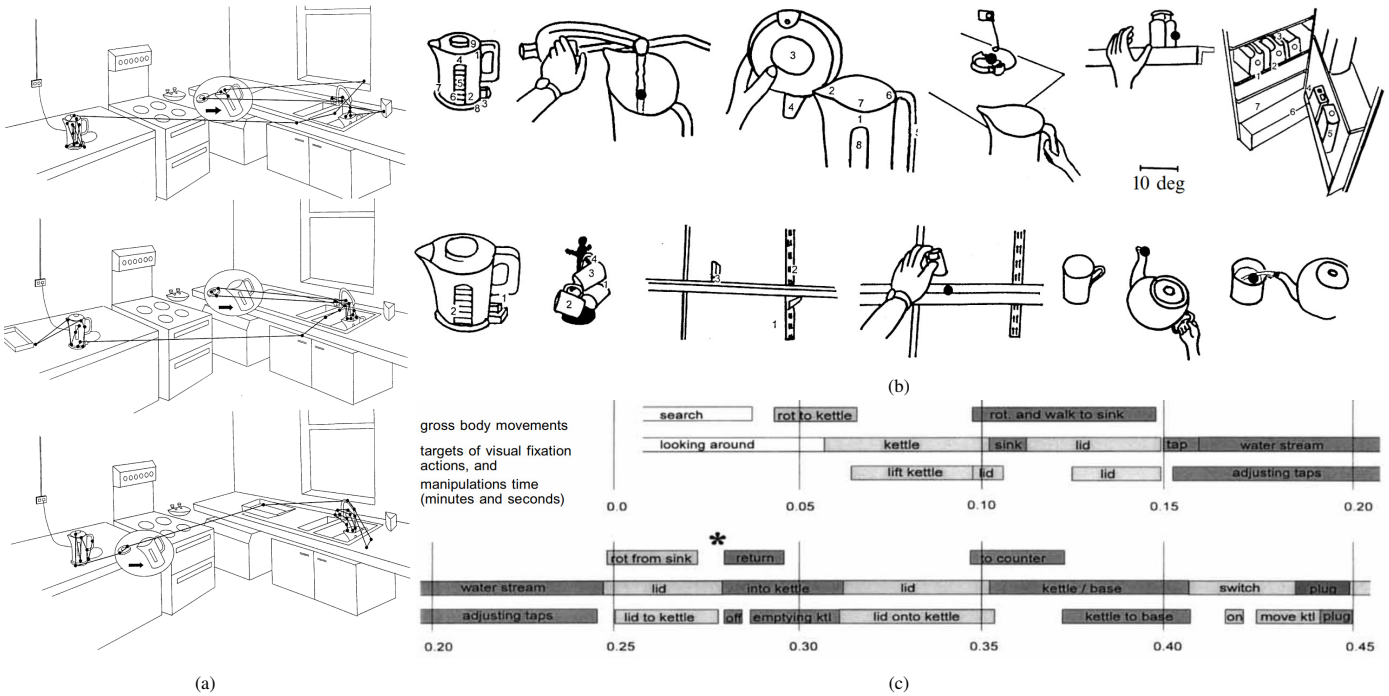


Figure 2: Even for a "simple" task as making a cup of tea, a person can make use of a single vision system to perform a variety of sub-tasks to achieve goals; images adapted from [3] with permission from the publisher. (a) Record of the fixations made by three different subjects performing the same task 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.

tion 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 five crucial axes of visual commonsense: **F**unctionality, **P**hysics, perceived **I**ntention, **C**ausality, and **U**tility (FPICU). 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. Ultimately, with the accumulated knowledge of the physical and social world, the decisions of a rational agent are utility-driven (Section 7). In a series of studies, we demonstrate that these five 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 8, 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. Vision: From Data-driven to Task-driven

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.

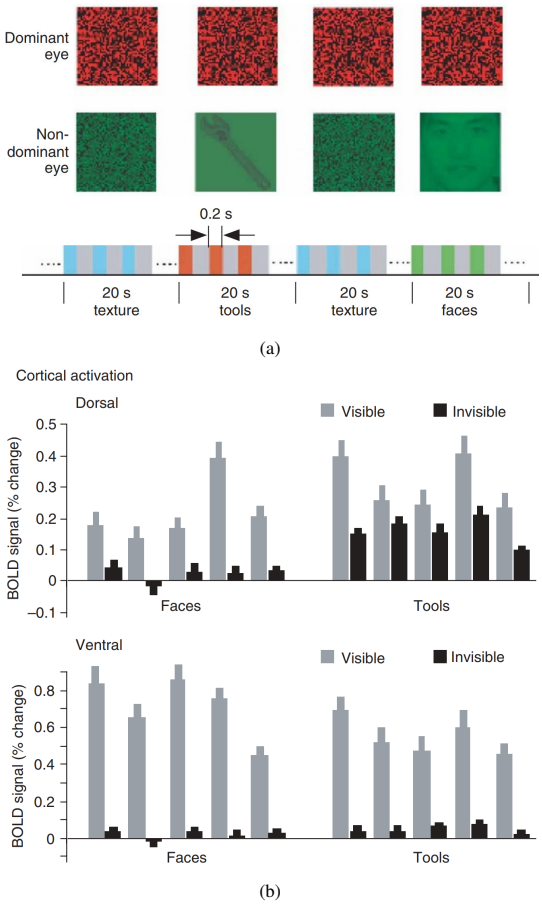


Figure 3: Cortical responses to invisible objects in the human dorsal and ventral pathways; images adapted from [5] with permission from the publisher. (a) Stimuli (tools and faces) and experimental procedures. (b) Both the dorsal and ventral areas responded to tools and faces. When stimuli were suppressed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

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 in the past decade and has achieved remarkable progress, including the recent success of DNNs [19, 20, 21].

Despite the fact that these approaches achieved a good performance on scene categorization in terms of the recognition accuracy on publicly available datasets, 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 were also found in face recognition [28].

grasp strategy	required functional capabilities	representation
	~center ~radius	superquadrics
	~center ~radius ~axis direction	generalized cylinder
	~center ~radius ~axis direction ~pulling direction	superquadrics + pulling direction
	orientation position of two planes width	two parallel planes (geometric model)
	center radius	cross-sectional shape (geometric model)
	position of points orientation	two contact positions (geometric model)

Figure 4: Different grasping strategies require various functional capabilities; images adapted from [7] with permission from the publisher.

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,

(a) (b) (c)

Figure 5: The experiment presented in [25], demonstrating a diagnostically driven bidirectional interplay between top-down and bottom-up information for how scenes are categorized at specific hierarchical levels; images adapted with permission from the publisher. (a) Given the same input image of a scene, subjects will show different gaze patterns if they were asked to categorize the scene at (b) a basic level (restaurant) or (c) a subordinate level (cafeteria), indicating a task-driven nature of scene categorization.

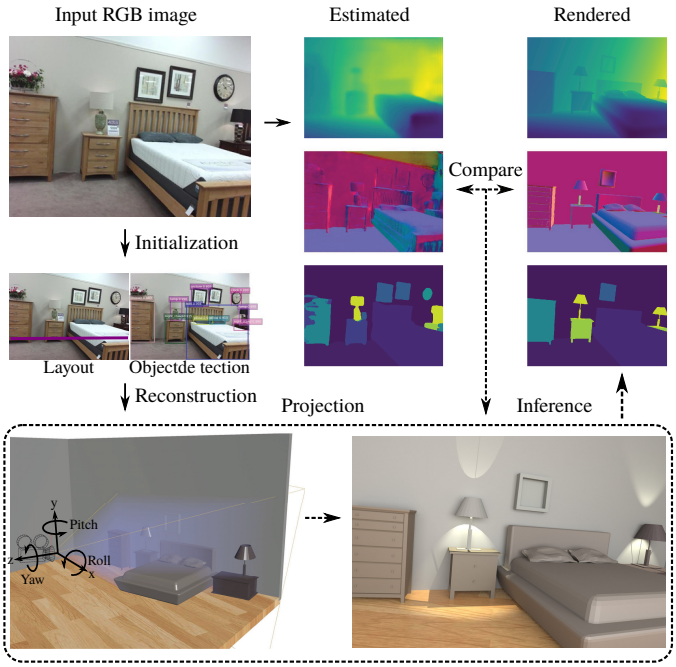


Figure 6: Illustration of the 3D indoor scene parsing and reconstruction in an analysis-by-synthesis fashion [36]. A 3D representation is initialized by individual vision tasks (*e.g.*, object detection, 2D layout estimation). A joint inference algorithm compares the differences between the rendered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

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 recognition 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 evidence 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-prone 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, implying 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 discoveries, 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 the 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

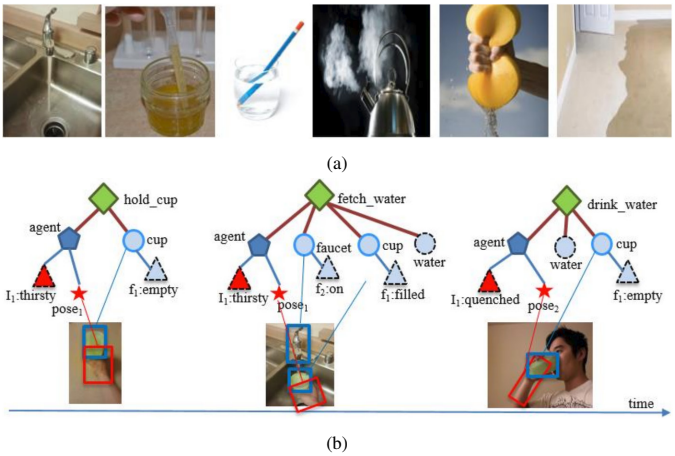


Figure 7: Water, and other fluid, play important roles in our activities but are hardly detectable in images. (a) Water causes minor appearance changes. (b) The 'dark' entities: water, fluents of the cup and faucet (triangle), and intent of human are shown in dashed nodes. The actions (diamonds) involves agents (pentagon) and cups (object in circles).

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 Scene Understanding with Human-like Commonsense

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 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 11.

To enable an artificial agent with similar 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, affordance, intents/goals, utilities, *etc*. These concepts could be divided into five 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 7. 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].

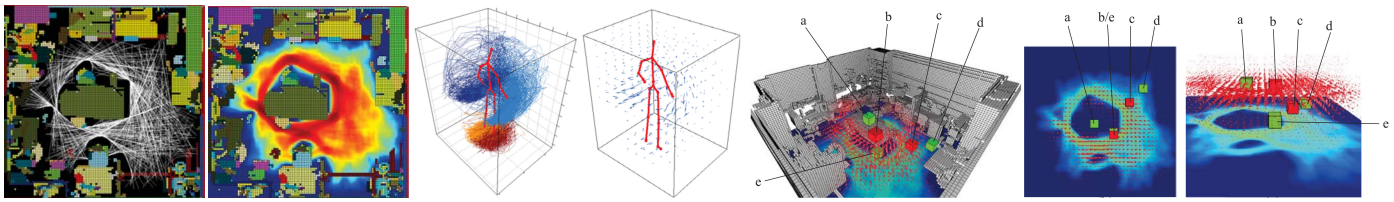
Fluents and perceived causality are different from the visual *attributes* [78, 79] of objects. The latter are permanent during the observation, *e.g.*, the gender of a person are attributes, not fluents. Some fluents are 'visible,' but many fluents are 'dark.' Human cognition has the innate capability (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 calling 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 understanding. 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, 86, 95], such as earthquake, wind/gust, and human activities. Therefore, any 3D scene interpretation or parsing (object and segmentation) must be physically plausible [94, 86, 95, 96, 36, 97]; see Figure 8. 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. 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.

By fMRI and neurophysiology experiments, researchers identified mirror neurons in the pre-motor cortical area that seem to encode actions through poses and interactions with objects and scenes [99]. Concepts in the human mind are not only represented by prototypes, *i.e*., exemplars in current vision



(a) (b) (c) (d)

Figure 8: 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.

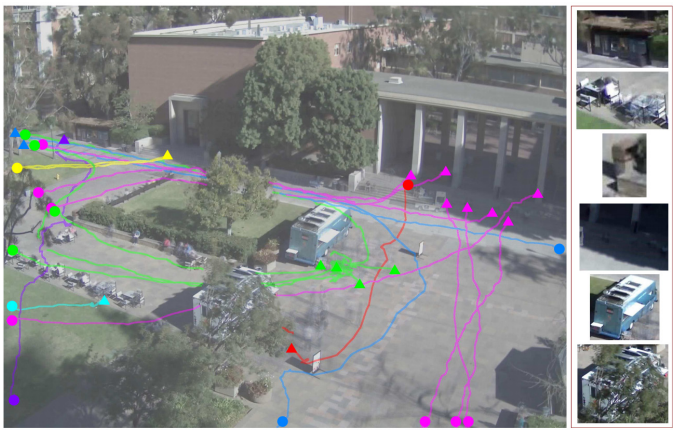


Figure 9: People's trajectories are color-coded by their shared goal destination [69]. The triangles denote destinations, and the dots denote start positions of the trajectories; *e.g.*, people may be heading toward the food-truck to buy food (green), or the vending machine to quench thirst (blue). Due to low resolution, poor lighting, and occlusions, objects at the destinations are very difficult to detect only based on their appearance and shape.

and machine learning approaches, but also by their functionality [77].

IV. Intents and Goals. Cognitive studies [100] 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 [101, 102].

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 spatiotemporal ranges. For instance, in Figure 9, 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 at a long distance. As 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 goals will be crucial for the following vision and cognition task: (1) Early event and trajectory prediction [103]. (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, 104] and affordance map [105, 106, 107] studied in the recent literature. (4) Understanding multi-way relations among a group of people and their functional roles [108, 109, 110]. (5) Understanding and inferring the mental states of agents [111, 112].

V. Utility and preference. Given an image or a video in which agents are 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 the rational choice theory; *i.e.*, a rational person's behavior, and decision-making are driven by maximizing their utility function. In the field of mechanism design in economics and game theory, this is related to the revelation principle, in which we assume each agent *truthfully* report their preference; see [113] 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 [114].

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 [115]. We argue such utility-driven learning could be more invariant than the traditional supervised training for computer vision and AI.

Summary. Despite their apparent differences at first glance, these five 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 and AI 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 such prior knowledge in these domains.

We argue that the incorporation of these five key elements will advance a vision and AI system in at least three aspects: (1) Generalization. As a higher-level representation, FPICU 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 FPICU encodes essential prior knowledge for understanding the environment, events, and behavior of agents. As FPICU is more invariant than appearance or geometric features, the leaning of FPICU, which is more consistent and noise-free across different domains and data sources, is possible even without "big data." (c) Bidirectional inference. Inference with FPICU 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 five key elements in greater detail.

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 [117, 118, 119]. People have innate assumptions about causes, and causal reasoning can be activated almost automatically and irresistibly [120, 121]. In our opinion, causality is the pillar for the other four topics (physics, functionality, intention, and utility). 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 learning causality in computer vision in AI.

3.1. Human Causal Perception and Reasoning

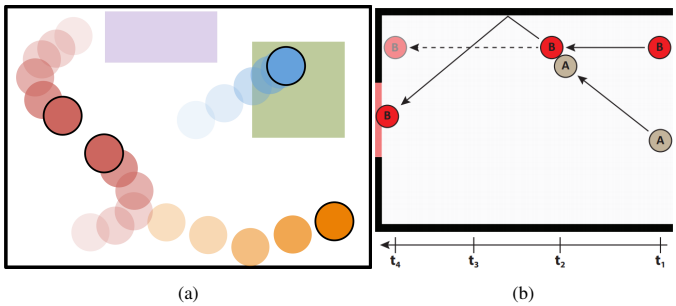
Humans reason about causal relationships through high-level cognitive reasoning. But can we "see" causality directly from vision, just as we see color and depth? In a series of behavioral experiments [122], Scholl's group showed the human visual system can *perceive* causal history by visual commonsense 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 [123].

Figure 10: Examples of some of Michotte's basic demonstrations of perceptual causality; images adapted from [116] with permission from the publisher. Perception of two objects, A and B (here shown as red and green circles). (a) The launching effect. (b) The entraining effect, wherein A seems to carry B along with it. (c) The launching effect is destroyed by adding a temporal gap between A's and B's motions. (d) The triggering effect, wherein B's motion is seen as autonomous, despite still being caused by A. (e) The launching effect is also destroyed by adding a spatial gap between A's final position and B's initial position. (f) The tool effect, wherein intermediate item (gray circle) seems merely a tool by which A causes the entire motion sequence.

Early psychological work focused on an associative mechanism as the basis for human causal learning and reasoning [124]. During this time, the Rescorla-Wagner model was used to explain how humans (and animals) build expectations using the co-occurrence of perceptual stimuli [125]. However, more recent studies have shown human causal learning is a rational Bayesian process [123, 126, 127] that involves the acquisition of *abstract* causal structure [128, 129] and strength values for cause-effect relationships [130].

The perception of causality is first systematically studied by psychologist Michotte [76] in the context of one billiard ball (A) hitting the other (B); see Figure 10. 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 [116] 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 mo-



(a) (b)

Figure 11: (a) 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. (b) 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 ball's actual movement paths. The dashed line indicates how Ball B would have moved if Ball A had not been present in the scene.

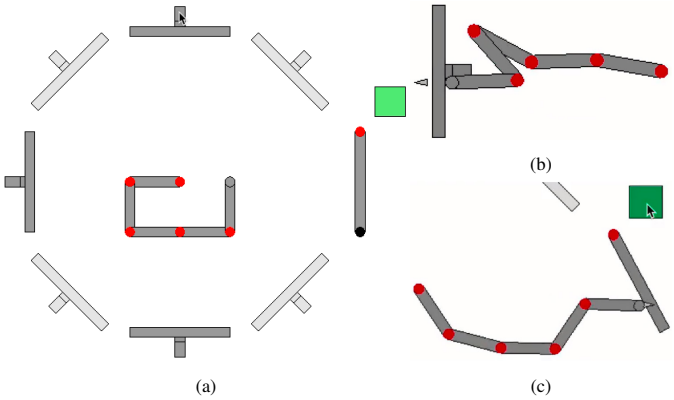


Figure 12: The OpenLock task presented in [129]. (a) Starting configuration of a 3-lever trial. All levers begin pulled towards the robot arm, whose base is anchored to the center of the display. The arm interacts with levers by either *pushing* outward or *pulling* inward. This is achieved by clicking 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.

tions. 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.

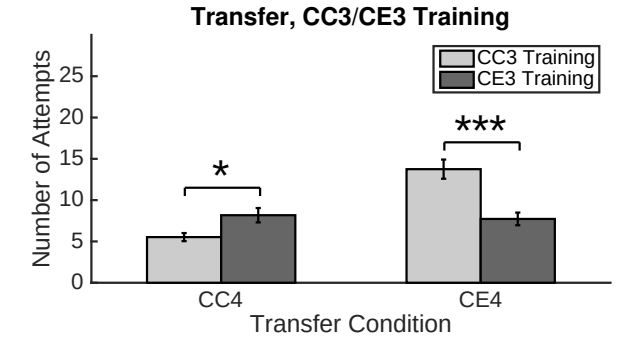
3. Richness. Even the interaction of two balls can support a variety of causal interactions. For example, if Ball B moves with a speed *faster* (vs. the same) than A, then the perception would not be that A "triggers" B's motion. Perceptual causality also includes "entraining," which is superficially identical to launching, except that A *continues* to move along with B once they make contact.

Recent cognitive science studies [131] 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 strong evidence of early visual processing of that stimuli. For example, it is well known that the perception of color can induce retinotopic adaption [132]. Strikingly, recent evidence revealed that there is a 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 [133] 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.

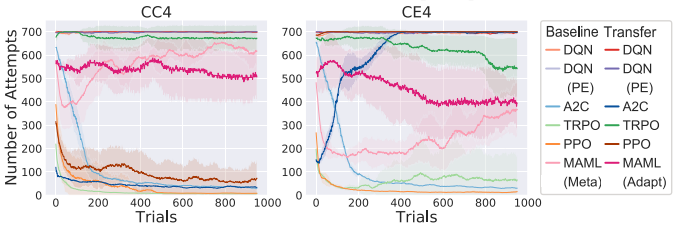
The importance of causal perception is beyond placing labels on different causal events. One unique function of causality is the support of counterfactual reasoning. Observers recruit the capacity of counterfactual reasoning to interpret visual events. In other words, interpretation is not based on what is observed, but on what would have happened but did not happen. In one study [134], 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 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.

3.2. Causal Transfer: Challenges for Machine Intelligence

Despite all the above evidence 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. However, learning causal concepts is of the utmost importance to agents expected to operate in observationally varying domains with common latent dynamics. To make this concrete, our world on Earth adheres to relatively constant environmental dynamics, *e.g.*, gravity is constant. Perhaps more importantly, our world is *designed* by other humans and largely adheres to common causal concepts: switches turn things off and on, knobs turn, doors open, *etc*. Even though objects in different settings appear different, their causal effect is constant because they all fit and cohere to a constant causal design. Thus, for agents expected to work in varying but human-designed environments, the ability to learn generalizable and transferable causal knowledge is crucial.



(a) Transfer trial results of human participants.



(b) Transfer trial results of RL agents.

Figure 13: Comparisons between human causal learners and typical RL agents [129]. Common Cause 4 (CC4) and Common Effect 4 (CE4) denote two transfer conditions used in Edmonds *et al.* [129]. (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.

Recent successes of systems such as deep Reinforcement Learning (RL) have showcased a broad range of applications [135, 136, 137, 138, 139], 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 [140, 141]. One approach to solving 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.* [129] 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 Figure 12. 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, latent 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 latent 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 (congruent) or different (incongruent) underlying causal structures to test whether the agent can generalize their acquired knowledge to more complex circumstances.

This task setting is unique and challenging for two major reasons: (1) transferring agents between rooms tests whether or not agents form *abstract* representations of the environment, and (2) transferring between 3- and 4-lever rooms examines how well agents are able to adapt causal knowledge to similar but different causal 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 13. Humans approached near-optimal performance and showed positive transfer to rooms with an additional lever in both congruent and incongruent conditions. 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.

3.3. Causality in Statistical Learning

Rubin laid the foundation for causal analysis in his seminal paper [143]; also see [144]. The formulation is commonly called the Rubin causal model. The key concept in the 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 [145], which is the conditional probability of assigning one treatment to the subject given the background

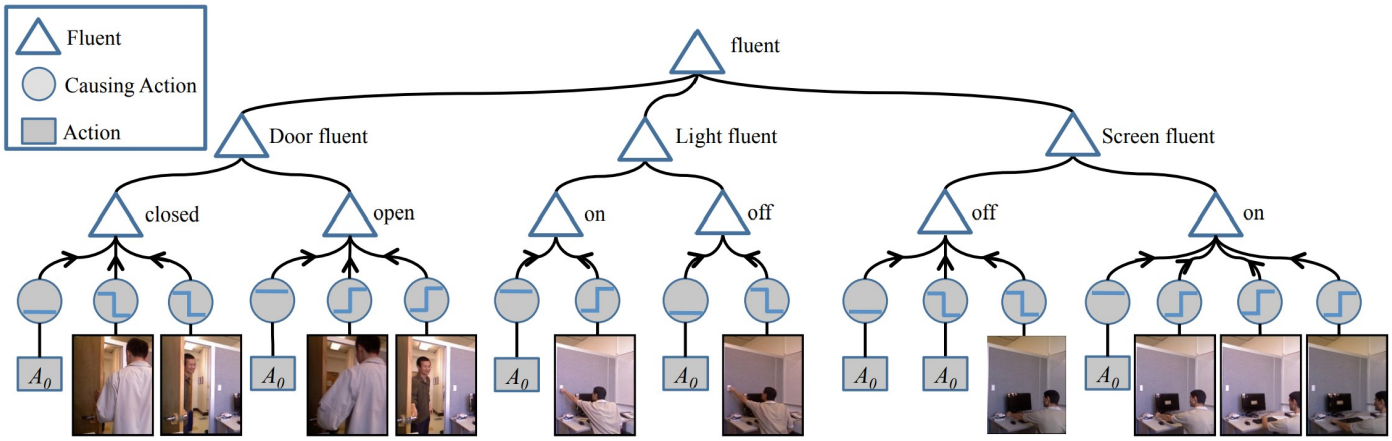


Figure 14: An example of perceptual causality in computer vision [142]. A Causal And-Or Graph for door status, light status, and screen status. Action A_0 represents non-action (a lack of state-changing agent action). Non-action is also used to explain the change of the monitor status to off when the screensaver activates. Arrows point from causes to effects, and undirected lines show deterministic definition.

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)) [146]. 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 [147]. 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 [147, 148, 149].

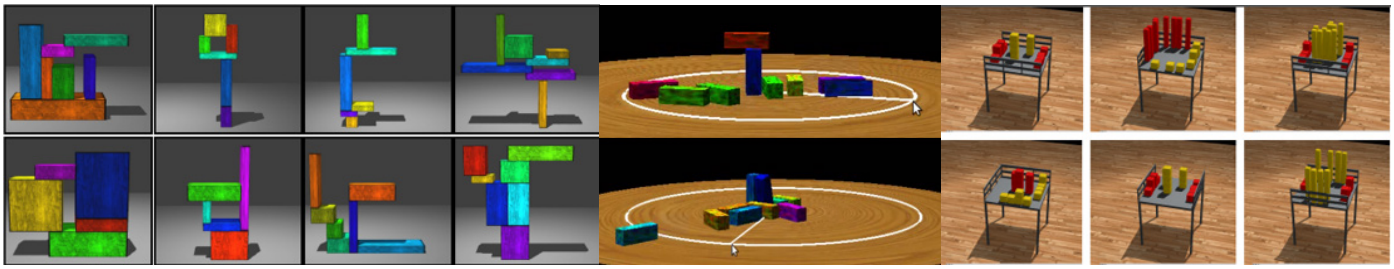
This framework offers a profound tool for fields where real-world interventions are difficult (if not impossible)—such as economics and epidemiology but lacks many properties necessary for human-like AI. Firstly, despite the attempts to learn the causal structure from observational data, most structure learning approaches cannot identify structure beyond a Markov equivalence class of possible structures [149]; 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 [150, 151]. 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 difficulty 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.

3.4. Causality in Computer Vision

The classical scientific setting for learning causality in clinical settings consists of Fisher's randomized controlled experiments [152]. 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 the human ability to perceive causal relationships from observations alone [116, 124, 125]. 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 and AI approaches should focus on how humans perceive causal relationships from observational data. Fire and Zhu [153, 142] proposed a method to learn causal relationships from image/video input; see an example in Figure 14. 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 [154] 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



(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 15: 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 physics judgment can be viewed as a form of probabilistic inference over the principles of Newtonian mechanics.

the information projection framework is capable of capturing the same judgments 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*. [155] 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 reason about the visibility fluent change jointly and track humans in videos. Experimental results demonstrate that with causal reasoning, they can recover complete trajectories of humans in complicated scenarios with frequent human interactions. Xiong *et al*. [156] also defined causality as a fluent change due to 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* [157]. The field of intuitive physics has been explored for several decades in cognitive science and recently reinvigorated by new techniques linked to AI.

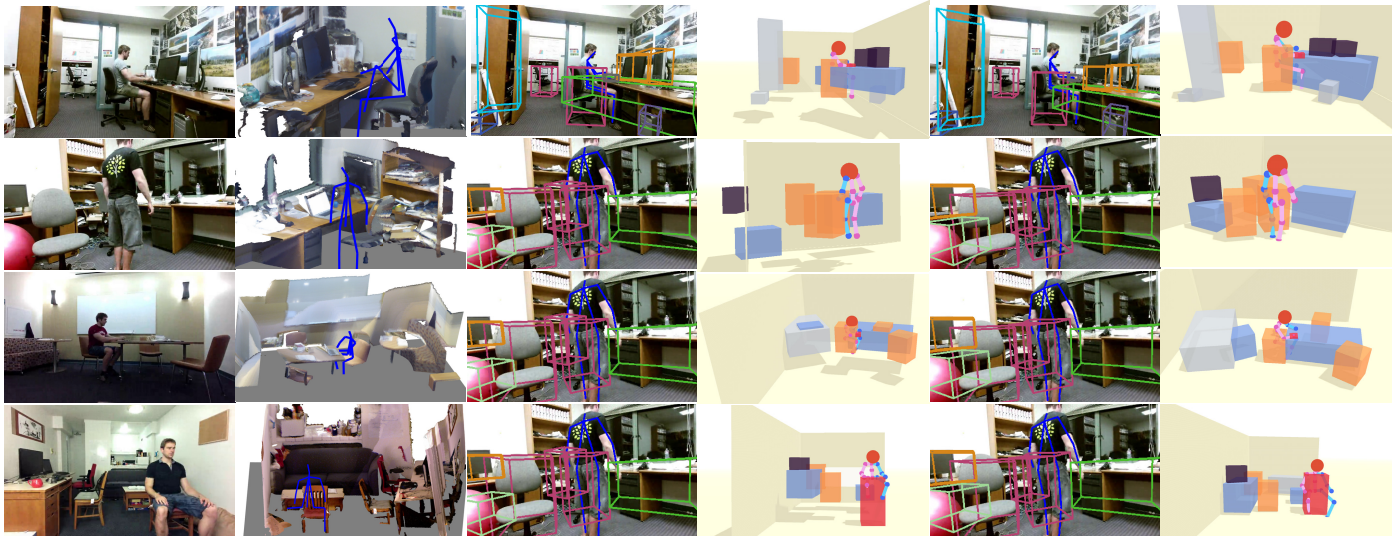
Surprisingly, humans develop physical intuitions at an early age [77], well before most other types of high-level reasoning, suggesting the importance of intuitive physics in comprehending and interacting with the physical world. The fact that physical understanding is rooted in visual processing also 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 explicitly reason about the expected continuation of a dynamic event based on a static image representing the situation at a single time point [158, 157, 159]. However, humans' intuitive understanding of physics is much more accurate, rich, and sophisticated than previously expected if *dynamics* and proper *context* were provided [160, 161, 162, 163, 164].

These findings are fundamentally different from prior work when it was systematically investigated the development of infants' physical knowledge [165, 166] in the 1950s. The reason to cause such a big difference is largely due to the fact that early work involves tasks beyond merely reasoning about physical knowledge, but also invoice other tasks [167, 168]. In address such difficulties, researches have developed alternative experimental approaches [169, 90, 170, 171] to study the development of infant's physical knowledge; most widely used approaches is the violation-of-expectation method, in which infants see two test events: an expected event, consistent with the expectation being shown in the experiment, and an unexpected event, violating the expectation. Such a series of studies have demonstrated strong evidence that humans, even young infants, possess expectations about various physical events [172, 173].

In a glance, humans can perceive whether a stack of dishes will topple, whether a branch will support a child's weight,



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

Figure 16: Scene parsing and reconstruction by integrating physics and human-object interactions [174]. (c) (d) Without adding physics, the parsed objects may flow in the air, resulting in an unnatural parsing. (e) (f) After adding physics, the parsed 3D scene becomes physically stable.

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 [175], 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 [176], 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 [177]; see Figure 15. Human intuitive 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 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 [178, 179] and sand [180]. In these studies, the experiments demonstrated that humans do not rely on simple qualitative heuristics to reason about fluid or

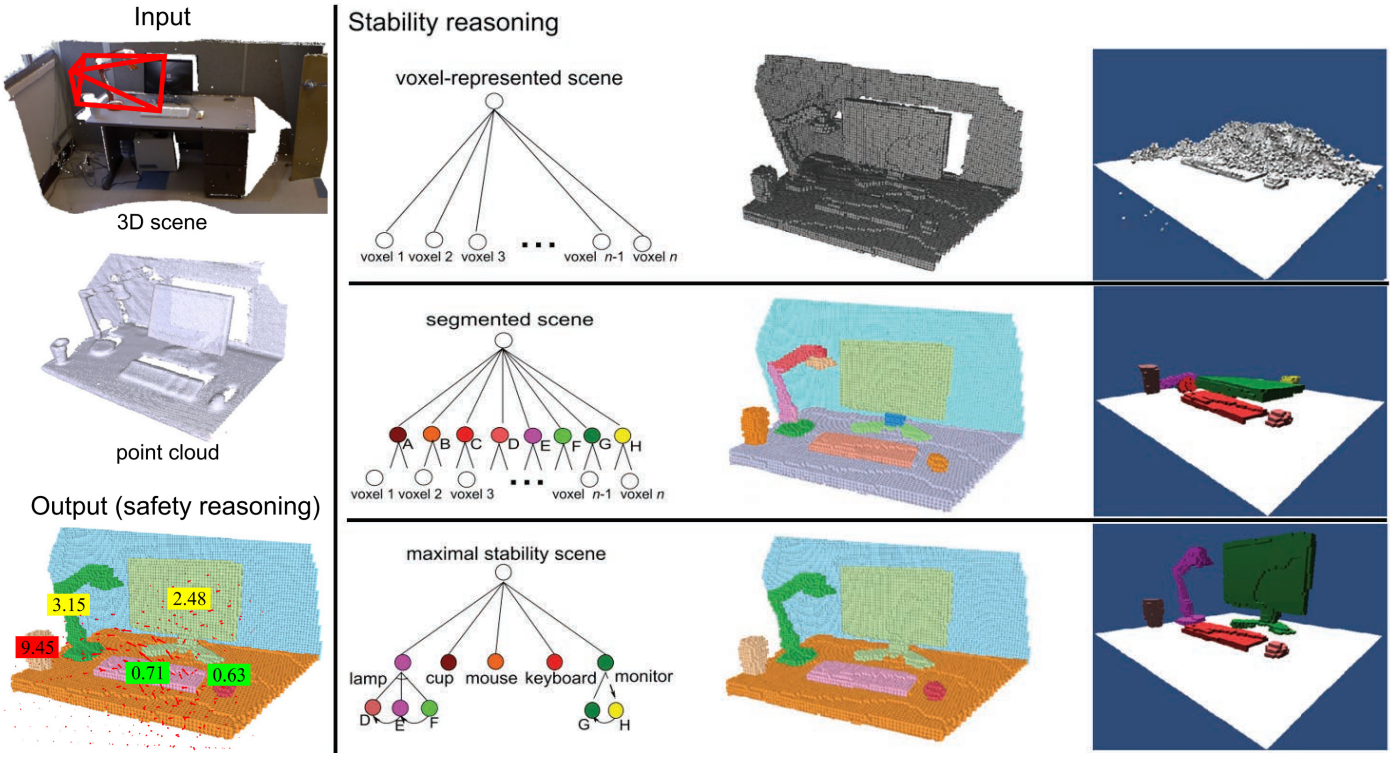


Figure 17: An example explicitly exploiting safety and stability in a 3D scene understanding task [95]. (a) Input: reconstructed 3D scene. Output: parsed and segmented 3D scene as stable objects. The numbers are unsafety scores for each object under the disturbance field (in red arrows) (b) Scene parsing graphs corresponding to 3 bottom-up processes: voxel-based representation (top), geometric preprocess including segmentation and volumetric completion (middle), and stability optimization (bottom).

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 of intuitive physics in psychology, see [181].

4.2. Physics-based Reasoning in Computer Vision

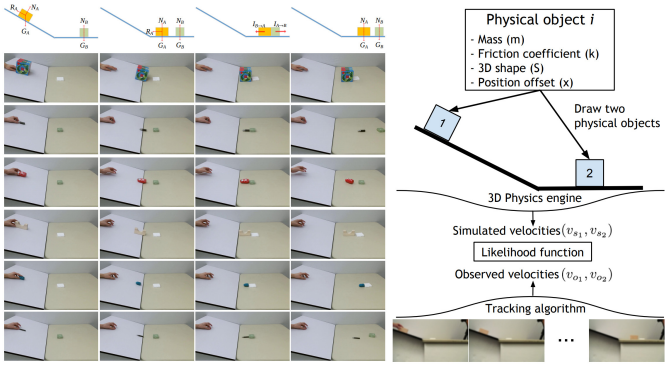
Classic computer vision focuses on appearance and geometric reasoning. Statistical modeling [182] aims to capture the "patterns generated by the world in any modality, with all their naturally occurring complexity and ambiguity, with the goal of reconstructing the processes, objects and events that produced them [183]." 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 [184, 185] and textons [186, 187] to form a primal sketch [188, 189], (2) a mid-level vision system to form 2.1D [190, 191, 192] and 2.5D [193] sketches, and (3) a high-level vision system in charge of full 3D [194, 195, 196]. In particular, he highlighted the importance of different levels of organization and the internal representation [197].

Alternatively, perceptual organization [198, 199] and Gestalt laws [200] 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 [201, 202] that are likely to be invariant over wide ranges of viewpoints [203], resulting in feature-based approaches [204, 84].

However, both approaches have well-known difficulties resolving the appearance [205] 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 16.

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 [95]. This line of work is mainly based on a simple but crucial observation 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



(a) Snapshots of datasets (b) Galileo model

Figure 18: Inferring the dynamics of the scenes [206]. (a) Snapshots of the dataset. (b) Overview of the Galileo model that estimates the physical properties of objects from visual inputs by incorporating the feedback of a physics engine in the loop.

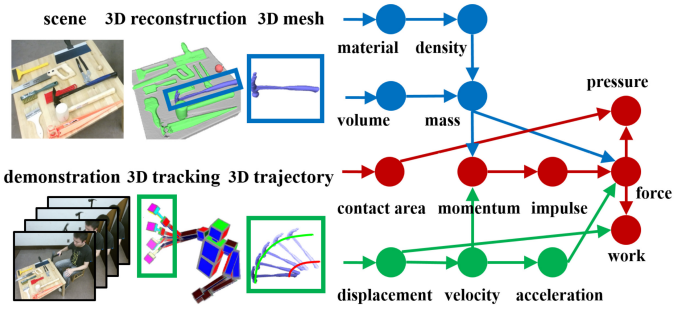


Figure 19: Thirteen physical concepts involved in tool-use and their compositional relations [221]. By parsing human demonstration, the physical concepts of material, volume, concept area, and displacement are estimated from 3D meshes of tool (blue), trajectories of tool-use (green) or jointly (red). The higher-level physical concepts can be further derived recursively.

deeper understanding of the 3D scenes beyond the observed pixels that could benefit various applications in robotics, virtual reality, and augmented reality.

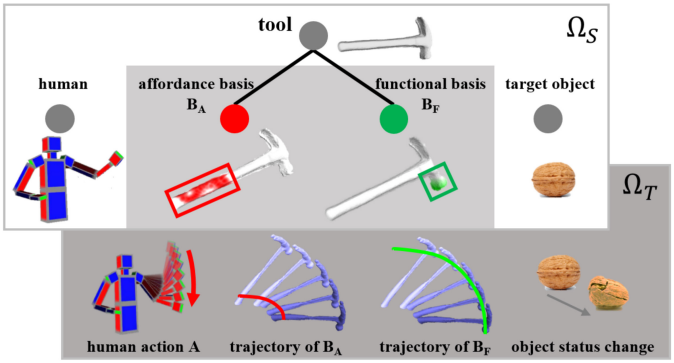
The ideas to incorporate physics to address vision problems could be traced back to Helmholtz and his argument for "unconscious inference" as part of the formation of visual impressions, whose function is to infer the probable causes of sensory input [207]. The very first formal solution in computer vision dates back to Roborts' solutions to completely parse and reconstruct 3D block world in 1963 [208], which inspired later researchers to realize the importance of the violation of physical laws for scene understanding [209] and the stability in generic robot manipulation tasks [210, 211].

Integrating physics for scene parsing and reconstructions was revisited in the 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 [212] scenes, where the algorithm infers the volumetric shapes of objects and relationships (occlusions and supporting relations) describing the 3D structure and mechanical configurations. In the next few years, other work [213, 214, 215, 216, 217, 106, 32, 218, 219, 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 [97, 174].

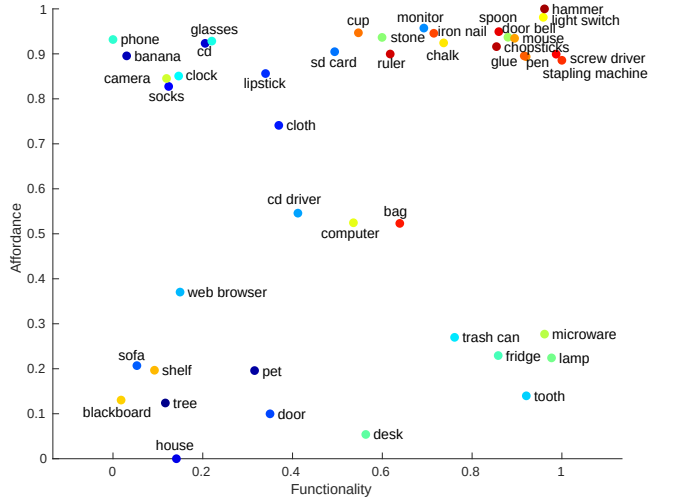
The aforementioned recent work mostly adopts simple physics cues; *i.e*., no or very limited physics-based simulation is applied. The first recent work that utilizes an actual physics simulator with modern computer vision methods was proposed by Zheng *et al*. in 2013 [94, 86, 95]. As shown in Figure 17, 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*. [220] by integrating an end-to-end trainable network and synthetic data.

Going beyond stability and supporting relations, Wu *et al*. [206] 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 18, 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 [222]. 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*. [221] calculated various physical concepts merely from a single example of tool-use (see Figure 19), 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*. [223] were able to infer the forces during human sitting behavior, resulting in an estimated force on various body mesh using a Finite Element Method (FEM); see Figure 35d.



(a) Funtional basis and affordance basis in a tool-use example.



(b) Examples of objects in the space spanned by functionality and affordance.

Figure 20: (a) The task-oriented representation of a hammer and its use in a task (crack a nut) in a joint spatiotemporal space. In this example, an object is decomposed into a functional basis and an affordance basis based on a given task. (b) The likelihood of a 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?" Vector graphics; zoom for details.

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*. [224, 225, 226] estimated contact forces and internal joint torques of human actions using a mass-spring system. Pham *et al*. [227] 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 [228, 229].

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

The perception of the environment inevitably leads to some course of action [230, 231]; 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" [232]; 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 [233], which is based on Gestalt theories and has a significant influence in changing the way we consider visual perception and scene understanding.

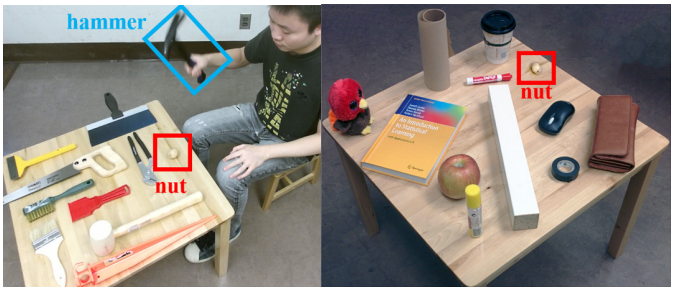
Similarly, a functional understanding of objects and scenes relates to identifying the possible set of tasks that can be performed with an object [234]. In contrast to affordances which are directly dependent on the actor, functionality is a permanent property of an object independent of the characteristics of the user; see an illustration in Figure 20. These two interweaving concepts are more invariant for object and scene understanding than their geometry and appearance dimensions. Specifically, we argue

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*., [235]). 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 computer vision and AI, 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 a complement to the previous reviews in data-driven approaches [236] and affordance tasks [237].

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 [238, 239]. Researchers have been using tool-use as the hallmark of the human intelligence to separate humans from non-human animals [240], until relatively recently that Dr. Goodall observed wild chimpanzees manufacture and use tools with regularity [241, 242, 243]. In addition to chimpanzees, studies have been reported on tool-uses by other species. For example, Santos *et al*. [244] 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*. [245] and Weir *et al*. [246] reported that New Caledonian crows can bend a piece of straight wire into a hook and successfully use it to lift a bucket containing food from a vertical pipe. More recent studies also found that New Caledonian crows behave optimistically after tool-using [247]—efforts cannot explain the optimistic; instead, they appear to enjoy, or be intrinsically motivated by, tool-use.



(a) Learning (b) Inference

Figure 21: Finding proper tool candidates in novel situations [221]. (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.

These discoveries suggest that some animals can and even be motivated to 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 interest and poses two major challenges in comparative cognition [248], which also hinders the reasoning ability in computer vision and AI systems.

First, why can some species come up with innovative solutions while others cannot when facing the same situations? See an example in Figure 21: 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 intraclass variance demonstrated in reasoning about the tool-use is extremely difficult to capture and resolve in the modern computer vision and AI systems. Without a consistent visual pattern, it is a long-tail problem as a visual recognition challenge; in fact, the very same object can serve for multiple functions depending on contexts and tasks. 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 possesses 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 possess such a reasoning capability? For example, New Caledonian crows are well-known for their propensity and dexterity at making and using tools. 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 [249]. 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 [233]

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 [250], 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 [233]. Based on the classic geometry-based "arch-learning" program [251], Winston *et al*. discussed the use of function-based descriptions of object categories [252]. 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 [253], Brady *et al*. proposed semantic net descriptions based on 2-D shapes together with a generalized structural description [254]. "Chair" and "Tool," exemplary categories researchers used for studies in functionality and affordance, were first systematically discussed with a computational method by Ho [255] and DiManzo *et al*. [256], respectively. Inspired by the functional aspect of the "chair" category in Minsky's book [257], the first work that uses a purely functional-based definition of an object category (*i.e.*, no explicit geometric or structural model) was proposed by Stark *et al*. [258]. These early ideas of integrating functionality and affordance with computer vision and AI systems have been modernized in the past decade; below, we review some representative work.

"Tool" is of particular interest 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*. [221] 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 22, 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 and AI 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*. [259] further learned the physical primitive decomposition for tool recognition and tower stability test. Using a more data-driven fashion, Fang *et al*. [260] extracted object affordance from labeled demonstration videos.

"Container" is ubiquitous in daily life and considered as a half-tool [261]. The study of containers can be traced back to a series of studies by Inhelder and Piaget in 1958 [262], in which they showed six-year-old children could still be confused by the complex phenomenon of pouring the liquid into

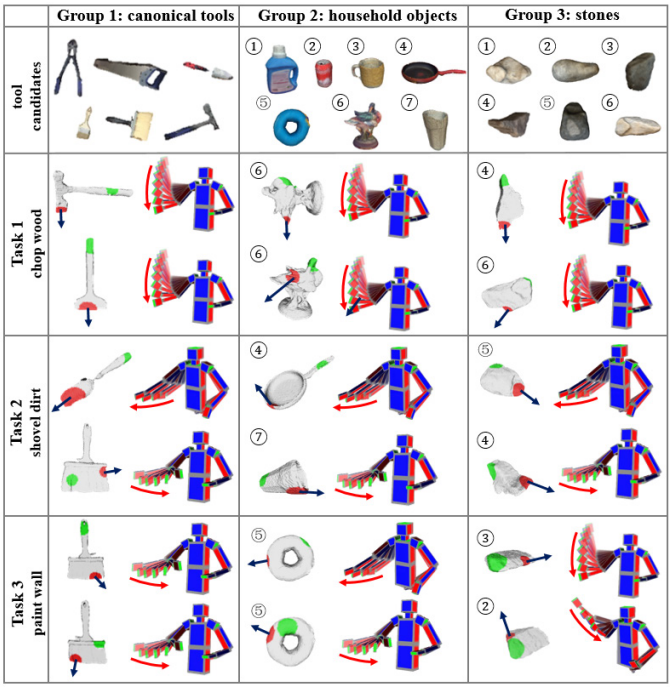


Figure 22: Given three tasks: chop wood, shovel dirt, and paint wall, the algorithm proposed by Zhu *et al*. [221] picks and ranks objects for each task among objects in three groups: (1) conventional tools, (2) household objects, and (3) stones, and output the imagined tool-use: affordance basis (the green spot to grasp with hand), functional basis (the red area applied to the target object), and the imagined action pose sequence.

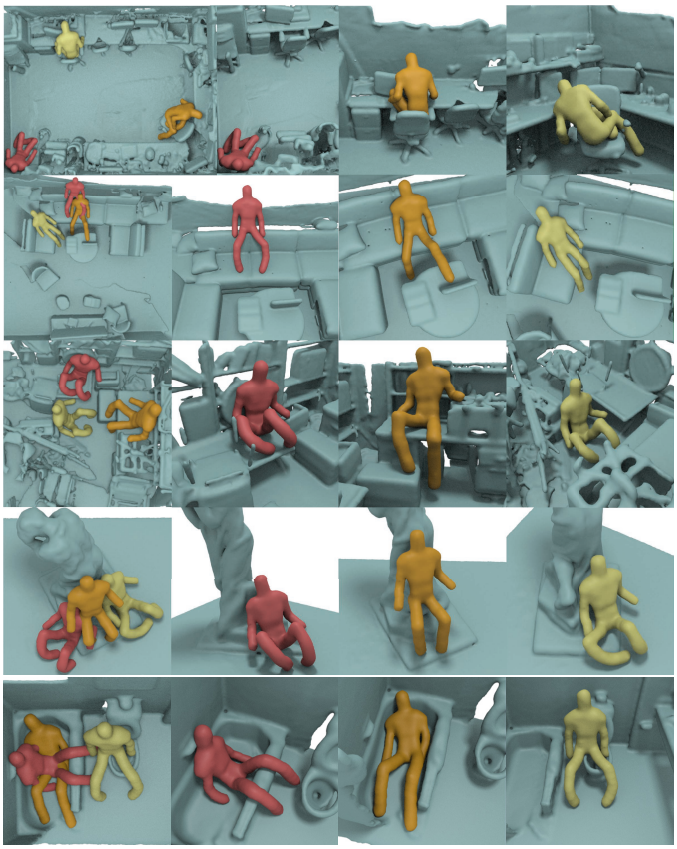


Figure 23: (a) Top 3 poses in various scenes for affordance (sitting) recognition [223]. Zoom-in views of the (b) best, (c) second, (d) third choice of the sitting poses. The top two rows are canonical scenarios, middle row is the cluttered scenario, and the bottom two rows are novel scenarios, which demonstrated a large generalization and transfer capability.

containers. Container and containment relations are of particular interest 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 [263] and support relations [264]). As early as 2.5 months old, infants can already understand containers and containment relations [265, 266, 267]. In AI community, researchers have been adopting commonsense reasoning [268, 269, 270] and qualitative representation/reasoning [271, 272] 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*. [273] 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, 161, 178, 179, 180, 181], which also enabled a few interesting directions and applications in computer vision, including reasoning about liquid transfer [274, 275], container and containment relation [276], and object tracking [277].

"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*. [105] 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*. [223] inferred the forces/pressures on various body parts while sitting on a chair; see Figure 23. 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. In this line of work, the methods all imagine the invisible human poses to help parse and understand the visible scene. 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 [233]. At the object-level, by learning human-object relations, Jiang *et al*. proposed methods to learn object arrangment [278] and object labelling [107] 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

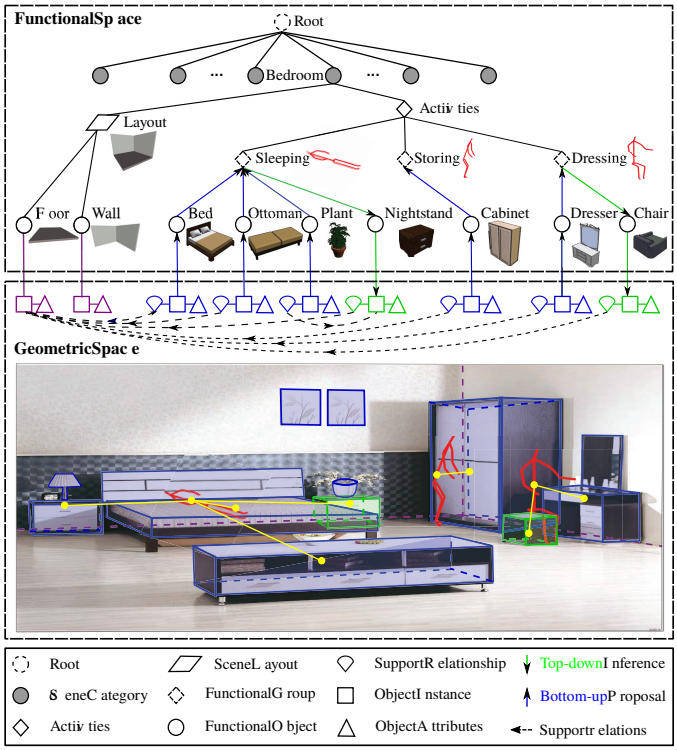


Figure 24: Task-centered representation of an indoor scene [36]. The functional space characterizes the hierarchical structure, and the geometric space encodes the spatial entities with contextual relations. The objects are grouped by the hidden activity groups, *i.e.*, by latent human context.

parse and reconstruct the 3D scene with a Holistic Scene Grammar (HSG). The HSG represents a functional task-centered representation of scenes. As shown in Figure 24, 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*. [96, 279] proposed human-centric representations to synthesize the 3D scenes with a simulation engine. As illustrated in Figure 25, they integrate human activities and functional grouping/supporting relations to sample more natural and reasonable activity spaces.

5.3. Mirroring: Causal-equivalent Functionality & Affordance

Unlike evaluating causality and physics, it is more difficult to evaluate the reasoning ability of a computer vision or AI system in terms of functionality and affordance. One effective way is to examine whether such information could endow more task capabilities to a learning system, *e.g.*, a robot. 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, requiring the algorithm to reason about the underlying mechanism instead of merely mimicking the motions. 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 26. This common problem is known as the "correspondence problem" [280] in Learing

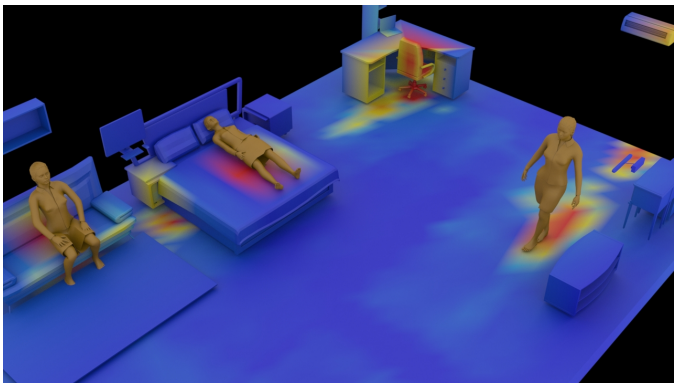


Figure 25: An example of the synthesized human-centric indoor scene (bedroom) with affordance heatmap generated by [96, 279]. The joint sampling of a scene is achieved by the alternative sampling of humans and objects according to the joint probability distribution.

from Demonstration (LfD); see more details in two previous surveys [281, 282].

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 causal-equivalent manipulation in terms of functionality and affordance—to imitate and replicate the task execution to achieve the same goal by reasoning about contact forces, instead of merely replicating the trajectory of the motion.

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 [227] often has limitations in handling self-occlusions and occlusions caused during manipulations. Other force-sensing devices such as strain gauge FlexForce [283] or the liquid-metal embedded elastomer sensor [284] 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*. [285] 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 demonstration is particularly important for 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

(a) Demonstration (b) Failure by pure imitation

Figure 26: (a) Given a successful human demonstration, (b) the robot may fail to accomplish the same task by imitating the human demonstration due to different embodiments. In this case, a two-finger gripper cannot firmly hold a hammer while swinging; the hammer slips, and the execution fails.

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 [285], the forces imposed to unlock the child-safety mechanisms of medicine bottles become observable. From these observations with latent force, Edmonds *et al*. [286] 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 [287].

However, the above work is still limited in the sense that the robot's actions are pre-defined, and the underlying structure of the task is not modeled. Recently, Liu *et al*. [288] 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 27. 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 over-imitation*: 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 of perceiving, planning, decision-making, and achieving goals in the environment. Crucially, it entails (1) *intentionality* [289] to represent the goal-state in the future and equifinal variability [290] to be able to

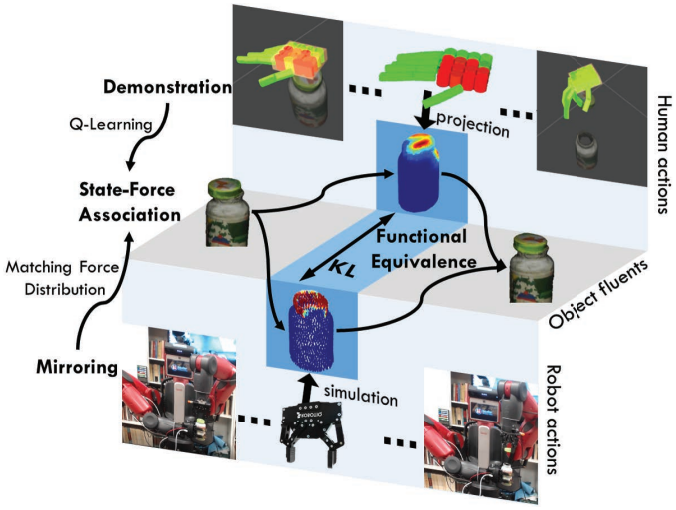


Figure 27: A robot mirrors human demonstrations with functional equivalence [288] by inferring the action that produces similar force, resulting in similar changes in the physical states. Q-Learning is applied to associate types of forces with the categories of the object state changes to produce human-object-interaction (*hoi*) units.

achieve the intended goal-state with different actions in various contexts, and (2) *rationality of actions* in relation to their goal [291] to produce the most efficient action and plan available. Perception and comprehension of intent enable humans to understand better 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 principle 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 six 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 of the building blocks for intention in computer vision and AI.

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 [292], 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 [293].

Evidence from developmental psychology shows that 6-month-olds see human activities as goal-directed behavior [294]. 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 [295, 296]. After their first birthdays, infants begin to understand that an actor may consider various action plans to pursue a goal and choose one to enact in intentional action based on reality [297]. 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 [298]. 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 [299]. Infants can also recover intentional relations at varying analysis levels, including concrete action goals, higher-order plans, and collaborative goals [300].

From infancy onward, we readily process action in intentional terms, despite the complexity of the behavioral stream we actually witness [293]. 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. The 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 [293]. The way to read intentions even creates species-unique forms of cultural learning and cognition [297]. From infants to complex social institutions, the world where we live is constituted from intentions of the agents present [301, 302, 297].

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 [303]; see Figure 28. 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

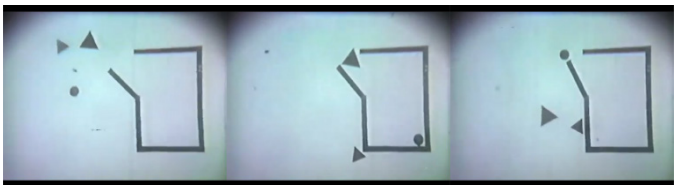


Figure 28: The seminal Heider-Simmel experiment [303]. Adults can perceive and attribute mental states merely from simple geometric shape motions.

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.

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 confounding variables by degrading the structural display while preserving the original dynamics [304]. They reported a similar number of anthropomorphic terms as in the original design, indicating the structure is not the critical information, which further strengthens 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 factors for the successful perception of the social relations and mental states. This phenomenon was later further studied by Bassili in a series of experiments [305].

Similar simulations of biologically meaningful motion sequences were produced by Dittrich and Lea [306], 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 are accorded with the Heider-Simmel experiment—motion dynamics play an important factor in the perception of intentional motion. Specifically, the intentionality appears stronger when the "wolf/lamb" path is more directly related to its target, and it is more salient when the speed difference is significant. Furthermore, they fail 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 demonstrate 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 arises: what is the underlying mechanism for the 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

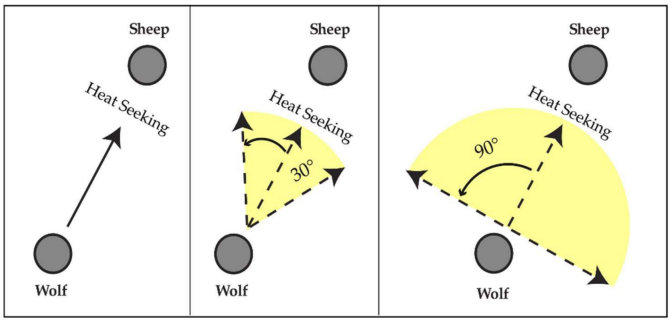


Figure 29: An illustration of the *chasing subtlety* manipulation in the "Don't-Get-Caught" experiment [309]. When the chasing subtlety is 0, the wolf always heads directly toward the (moving) sheep, in a "heat-seeking" manner. When the chasing subtlety is 30, the wolf is always heading in the general 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.

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 [307, 291, 308].

Guided by this principle, Gao *et al*. [309] has explored the psychophysics of chasing, one of the most salient and evolutionarily important types 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 29. 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 [311, 312]. 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 [294]. Recently, Liu *et al*. [308] performed five looking-time experiments with 3-month-old infants, in which infants viewed object-directed reaches that varied in efficiency (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 [101, 313, 111]. 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*. [101] 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 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 [314, 315, 316, 313].

Going from the symbolic input to real video input, Holtzen *et al*. [310] 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 30. 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*. [109] proposed a method to infer humans' intention to interact from motion trajectories; see Figure 31. A non-parametric exponential potential function is learned 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 about 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 9. 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

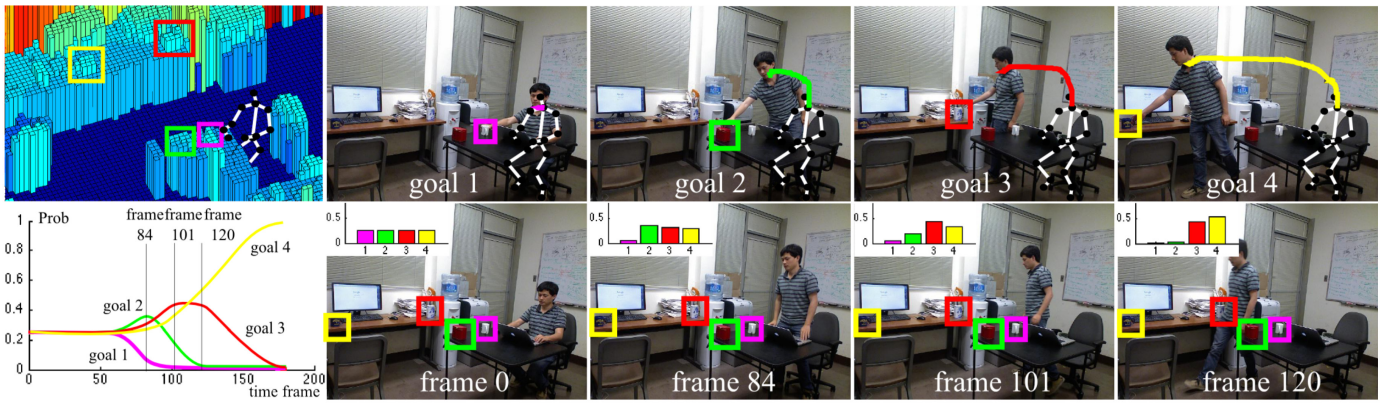


Figure 30: The plan inference task presented in [310]; seen from the perspective of an observing robot. The top left panel shows four different goals (target objects) in a 3D scene. The bottom left panel shows one outcome of the proposed method: the marginal probability of each terminal action over time. 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.

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 and agents' intents.

6.3. Beyond Action Prediction

Intention is related to action prediction in modern computer vision and AI systems [317], *much* more than predicting merely an action label; humans have a strong and early inclination to interpret actions in terms of intention as a long-term *social learning* of novel means and novel goals. From a philosophical view, Csibra *et al*. [100] contrasted three distinct mechanisms: (1) action-effect association, (2) simulation procedures, and (3) teleological reasoning. They concluded that action-effect association and simulation could 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 [318], and such simulation processes are at the heart of the development of intentional action interpretation [298]. In order to understand others' intentions, 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 [319] 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 [320]. Accumulating neurophysiological evidence support such simulations in the human brain, such as the mirror neurons [321], which has been linked to intention understanding by many studies [322, 99]. 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*., [323]), which poses challenges to the simulation theory.

To address social learning, teleological action interpretational system [324] takes a 'functional stance' for the computational representation of goal-directed action [100], where such teleological representations are generated by the aforementioned inferential "rationality principle" [325]. In fact, the very notion of 'action' implies a motor behavior 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

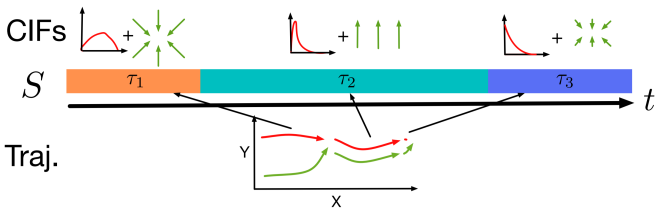


Figure 31: Inference of human interaction from motion trajectories [109]. The top demonstrates the change of conditional interactive field (CIF) in sub-interactions as the interaction proceeds, where the CIF models the expected relative motion pattern conditioned on the reference agent's motion. The bottom indicates the change of interactive behaviors in terms of motion trajectories. The colored bars in the middle depict the types of sub-interactions.

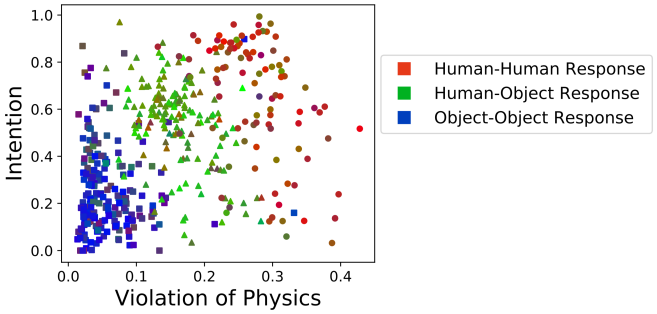


Figure 32: Constructed psychological space including HH animations with 100% animacy degree, HO animations, and OO animations [110]. Here, a stimulus is depicted by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. 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).

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.* [110] 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. 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 32. 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 [326]. 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 [327]. Social eye gaze functions also transcend cultural differences, forming a kind of universal language [328]. Computer vision and AI systems heavily rely on gazes as cues for intention prediction based on images and videos. For instance, Wei *et al.* [329] 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 in terms of hand-eye coordination under a planner represented by a grammar; see Figure 33.

Communicative gazes and gestures (*e.g.*, pointing) stand out for intention expression and perception in collaborative interactions. Humans need to recognize partners' communicative intentions to collaborate with others and survive in the world successfully. Human communication in mutualistic collaboration often involves agents informing recipients of things they believe will be useful or relevant to them. Tomasello *et al.* [331] 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 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. To model such a capability demonstrated in both chimpanzees and humans, Fan *et al.* [332] studied the problem of human communicative gaze dynamics—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 [330] studied various types of gaze communications in social activities from both atomic-level and event-level; see Figure 34. A spatiotemporal graph network is proposed to explicitly represent the diverse interactions in the social scenes and infer atomic-level gaze communications.

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 equip such an ability with robots to act naturally and properly subject to "social affordance", which represents action possibilities following basic social norms. Trick *et al.* [333] proposed an approach for multimodal intention recognition considering four modalities, including speech, gestures, gaze directions, and scene objects,

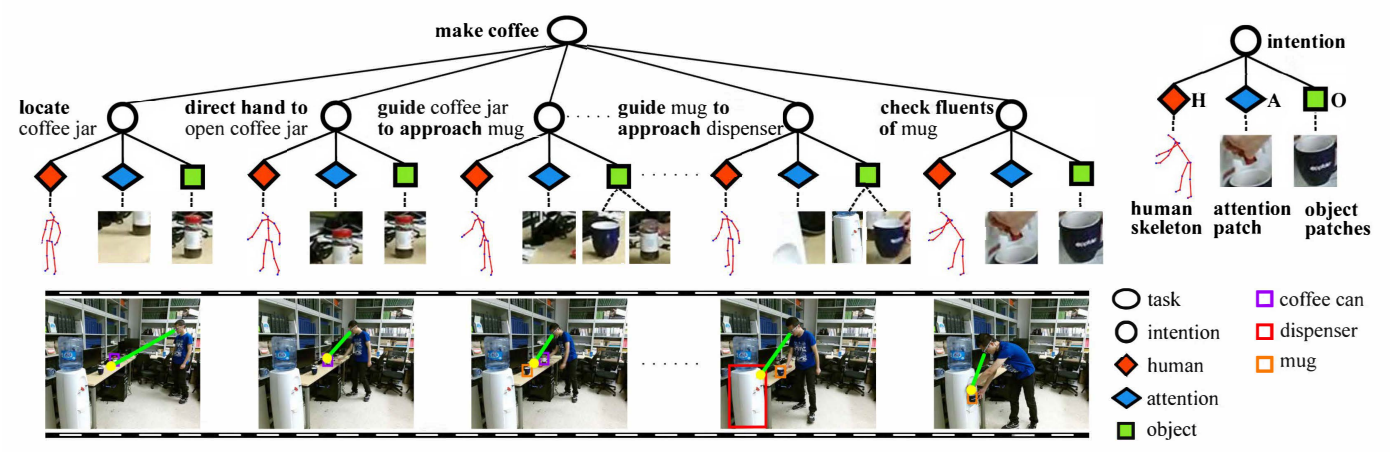


Figure 33: Model a task as sequential intentions in terms of hand-eye coordination by Human-Attention-Object (HAO) graph [329]. Here, an intention is represented as inverse planning, in which human pose, attention, and object provide contexts to afford the inference about an agent's intention.

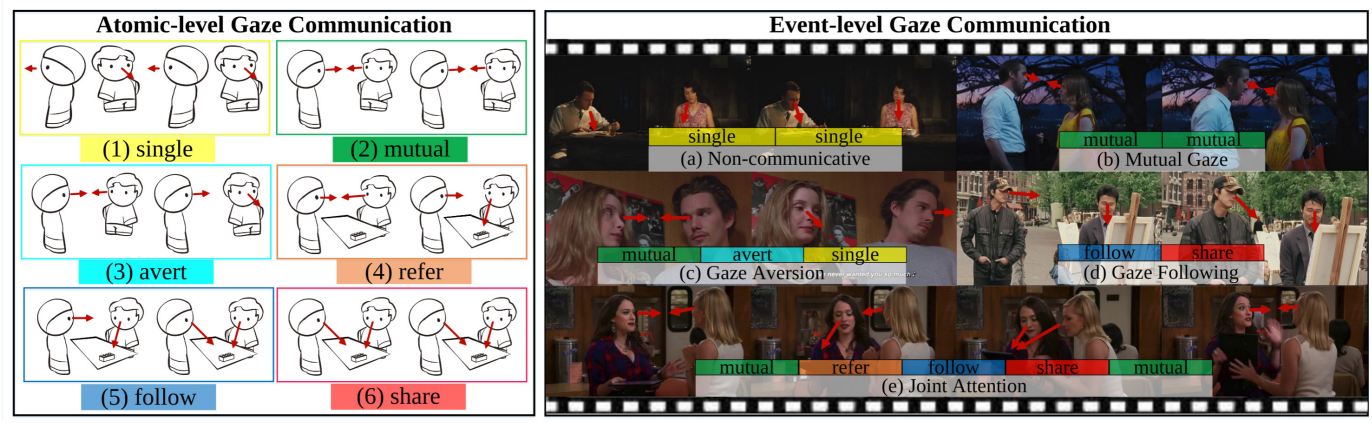


Figure 34: Human gaze communication dynamics in two hierarchical levels [330]: (1) atomic-level gaze communication describes the fine-grained structures in human gaze interactions, and (2) event-level gaze communication refers to long-term social communication events temporally composed of atomic-level gaze communications.

focusing on uncertainty reduction through classifier fusion. Shu *et al*. [334] presents a generative model for robot learning of social affordance from human activity videos. By discovering critical steps (*i.e*., latent sub-goals) in 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 [335], 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. Learning Utility – The Preference of Choices

Rooted from the field of philosophy, economics, and game theory, the concept of utility serves as one of the most basic principles of modern decision theory: an agent makes rational decisions/choices based on what they believe and what they want to maximize the agent's expected utility, known as the principle of maximum expected utility. We argue that the majority of the observational signals we encounter in daily life are largely driven by this simple yet powerful principle—an invisible "dark" force that governs the underlying mechanism explicitly or implicitly of human behaviors. Thus, studying utility could afford a computer vision or AI system with a deeper understanding of the observations, thereby achieving better generalization.

By classic definitions, the utility that the decision-maker obtains from selecting a specific choice is measured by a utility function, a mathematical formulation that ranks the preferences of the individual such that $U(a) > U(b)$, where the choice a is preferred over the choice b. It is important to note that the existence of a utility function that describes an agent's preference behavior does not necessarily mean that the agent is *explicitly* maximizing that utility function in their own deliberations. By observing a rational agent's preferences, however, an observer could construct the utility function that represents what the agent is actually trying to achieve, even if the agent

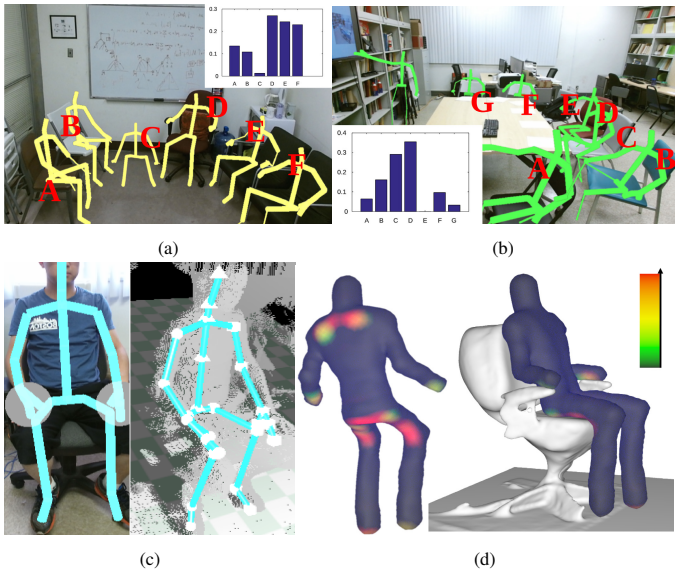


Figure 35: Examples of sitting activities in (a) an office and (b) a meeting room [223]. In addition to geometry and appearance, people also consider other important factors, including comfortability, reaching cost, and social goals when choosing a chair. The histograms indicate human preferences for different candidate chairs. Based on these observations, one can infer human utility during sitting activities from videos [223]. (c) 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. (d) Using FEM simulation, the forces are estimated at each vertex of the FEM mesh.

does not know it [336]. It is also worth noting that the utility theory is a *positive* theory that seeks to explain the individuals' *observed* behavior and choices, which is different from a *normative* theory that indicates what people *should* behave; such a distinction is crucial in the discipline of economics and for us to devise algorithms and systems to interpret the observational signals.

Although Jeremy Bentham [114] is often regarded as the first to systematically study utilitarianism—the philosophy concept later borrowed to economics and game theory, the core insight motivating the theory occurred much earlier; *e.g.*, Francis Hutcheson on action choice [337]. In the field of philosophy, Utilitarianism is considered as a normative ethical theory that places the locus of right and wrong solely on the outcomes (consequences) of choosing one action/policy over other actions/policies. As such, it moves beyond the scope of one's own interests and takes into account the interests of others [337, 338]. The term has been adapted in the field of economics; a utility function represents a consumer's preference ordering over a choice set, which is now devoid of its original meaning as a measurement of the pleasure or satisfaction.

Formally, the core idea behind utility theory is straightforward: every possible action or state within a given model can be described with a single, uniform value. This value, usually referred to as *utility*, describes the usefulness of that action within the given context. Note that the concept of the *utility* is not the same as the concept of *value*: utility measures how much we desire something in a more subjective and context-dependent perspective, whereas value is a measurable quantity (*e.g.*, price), which tends to be more objective. To demonstrate the usefulness of adopting the concept of utility into a computer vision and AI system, we briefly review three recent case studies in computer vision, robotics, and linguistic, using a utility-driven learning approach.

As shown in Figure 35, by observing the choices people make in videos (particularly in selecting a chair in which to sit), a computer vision system [223] 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.

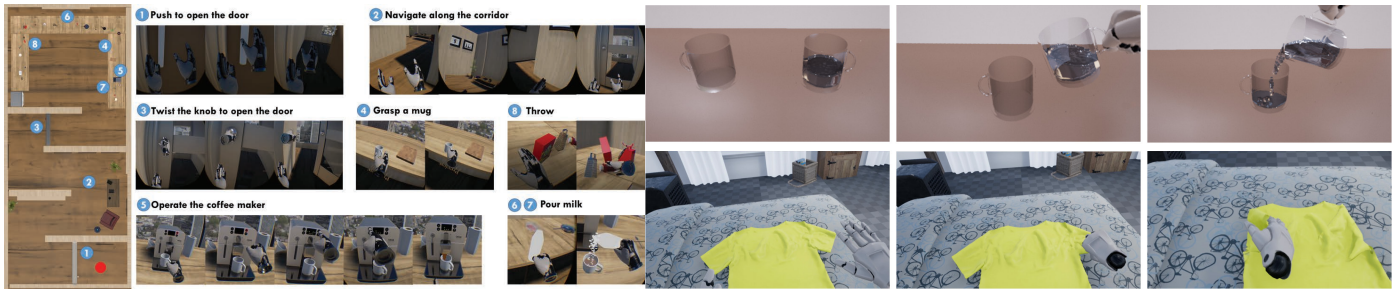
Similarly, Shukla *et al*. [340] adopted the idea of learning human utility in order to learn from human demonstrations for a robotics task. A proof of concept work shows a pipeline in which the agent learns the *external* utility of humans and plans with the learned utility functions for a cloth-folding task. Specifically, under the assumption that the utility of goal states is better than initial states, this work learns the *external* utility of humans by ranking the pairs of states extracted from images.

The rational principle has also been studied in the field of linguistics and philosophy, notably from Grice's influence work on the theory of implicature [341]. 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 [342]. Such a goal-directed view of language production has led to a few interesting language games [343, 344, 345, 346, 347, 348], the development of engineering systems for natural language generation [349], and a vocabulary for formal descriptions of pragmatic phenomena in the field of game-theory [350, 351]. More recently, by assuming the communications between the agents to be helpful yet parsimonious, the 'Rational Speech Act' [352, 342] model has demonstrated promising results in solving some challenging referential games.

8. Summary and Discussions

Robots are mechanically capable of performing a wide range of human activities; but in practice, they do very little for us. Fundamentally, robots still lack physical and social commonsense, and this limitation inhibits their capacity to aid in our daily lives. In this article, we reviewed five crucial aspects as the building blocks of commonsense: functionality, physics, intention, causality, and utility (FPICU). We argue these cognitive aspects are the foundation for constructing a cognitive architecture of future computer vision and AI. The position and opinions provided in this article do not intend to serve as *the* solution for the future of cognitive AI; 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:

- **Physically-Realistic VR/MR Platform: From Big-Data to Big-Tasks.** Since FPICU is "dark"—often does not directly appear in any pixels, it is difficult to evaluate FPICU in traditional terms. Here, we argue that the ultimate standard for



(a) Various task executions in VRGym

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

Figure 36: VRGym [339], 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 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.

validating the effectiveness of FPICU 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 and physics-based simulation, we begin to explore this possibility at a large scale; see Section 8.1.

- **Social System: Emergence of Language, Communication, and Morality.** While FPICU captures the core components to model a single agent, interaction among troops of agents, either in collaborative or competitive situations [357], is still a challenging problem. In most cases, the algorithms designed for a single agent would be complicated to generalize to an Multiple-Agent Systems (MAS) setting [353, 354, 355]. We provide a brief review of three related topics in Section 8.2.
- **Measuring the Limits of Intelligence System: IQ tests.** Studying FPICU opens a new direction of analogy and relational reasoning [358]. Apart from the four-term analogy, or proportional analogy, John C. Raven [359] proposed the Raven's Prograssive Matrices Test (RPM) in the image domain. The RAVEN dataset [360] 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 8.3.

8.1. Physically-Realistic VR/MR Platform: From Big-Data to Big-Tasks

A hallmark of machine intelligence is the capability to rapidly adapt to new tasks and "achieve goals in a wide range of environments" [361]. 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 [362, 363, 279, 364], including MINOR [365], HoME [366], Gibson [367], House3D [368], AI-THOR [369], VirtualHome [370], VRGym [339] (see Figure 36), VRKitchen [371], *etc*. Similarly, AirSim [372] 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. In these realistic virtual environments, one can evaluate any AI methods or systems in a much more holistic perspective. By such an evaluation, whether a method or a system is intelligent is no longer measured by achieving a good performance in a single task; rather, it demands to evaluate across various tasks: the perception of the environments, the planning of the actions, the predictions of other agents' behaviors, and the ability to rapidly adapt learned knowledge to new environments for new tasks.

To afford such a task-driven evaluation, physics-based simulation for multi-material multi-physics phenomena (see Figure 37) will play a central role. 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 [373]. Here, we provide a brief review of the recent physics-based simulation methods, in particular, Material Point Method (MPM).

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 the simulation of various solids and fluids in a 3D virtual environment, since the pioneering work of Terzopoulos *et al*. [374, 375] for solids and Foster *et al*. [376] 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 compan-

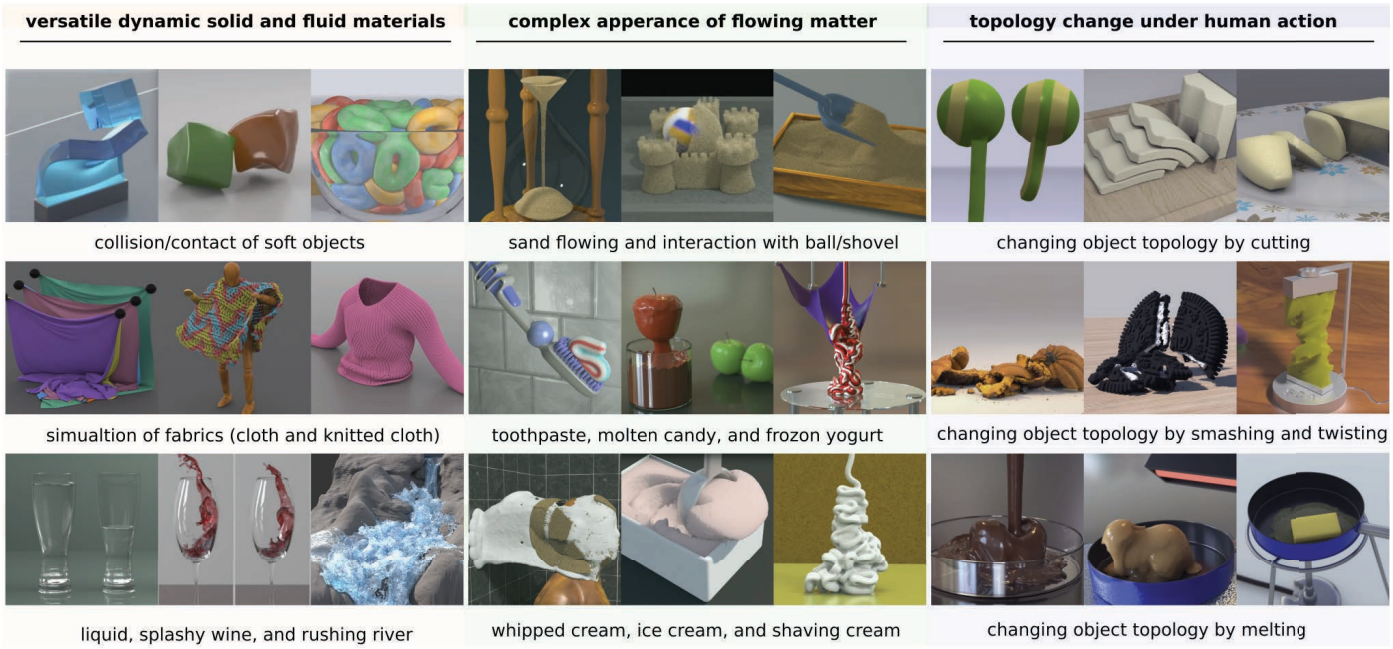


Figure 37: Diverse physical phenomena simulated using Material Point Method (MPM).

ion 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 [377, 378]. 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 [379, 380, 381], 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 [382, 383] or topologically changing events [384].
3. Lagrangian mesh-free methods, such as Smoothed Particle Hydrodynamics (SPH) [385] and Reproducing Kernel Particle Method (RKPM) [386]. These methods allow arbitrary deformation but require expensive operations such as neighborhood search [387]. 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) [388], 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 [389, 390] from computational 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 [391, 392]. In the field of visual computing, the existing work include snow [393, 394], foam [395, 396, 397], sand [398, 399], rigid body [400], fracture [401, 402], cloth [403], hair [404], water [405], and solid-fluid mixtures [406, 407, 408]. 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 [409], material failure [410, 411], hyper-velocity impact [412, 413], explosion [414], fluid-structure interaction [415, 416], biomechanics [417], geomechanics [418], 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 [419]. 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 [420, 421], Convected Particle Domain Interpolation (CPDI) method [422] and Dual Domain Material Point (DDMP) [423].

8.2. Social System: Emergence of Language, Communication, and Morality

Being able to communicate and collaborate with other agents is a crucial component of AI. In classic AI, a multi-agent communication strategy is modeled using a predefined rule-based system (*e.g.*, adaptive learning of communication strategies in MAS [357]). To scale up from rule-based systems, decentralized partially observable Markov decision processes are devised to model multi-agent interaction with communication as a special type of action among agents [424, 425]. As the success of RL in single-agent games [426], generalizing Q-learning [427, 355] and actor-critic [353, 428] based methods from single-agent to MAS has been a booming topic in recent years.

The emergence of language is also a fruitful topic in multi-agent decentralized collaborations. By modeling communication as a particular type of action, recent research [354, 432, 433] has shown that agents can learn how to communicate with continuous signals only decipherable within a group. The emergence of more realistic communication protocols using discrete messages has been explored with various types of communication games [434, 435, 436, 437], in which agents need to process visual signals and ground discrete tokens to attributes or semantics of the images to form effective protocols. By letting groups of agents play communication games spontaneously, several linguistic phenomena in emergent communication and language have been studied [438, 439, 440].

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 every day under the guidance of those principles, and the trade-offs often represent innate conflicts posited on the principles centered around morality [441, 442]. Moral judgment is extremely complicated due to the variability in standards among different individuals, social groups, cultures, and even forms of violation of ethical rules. For example, two distinct societies could hold opposite views on preferential treatment of kin: corruption or moral obligation [443], and the same principle might be viewed differently in two social groups with distinct cultures [444]. Even within the same social group, different individuals might have different standards on the same moral incident or principle [445, 446, 447]. Many works proposed theoretical accounts for categorizing welfare used in morality calculus, including "base goods" and "primary goods" [448, 449], "moral foundations" [450], and ability of value judgment from infants' points of view [451]. 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 [443, 452, 453].

8.3. Measuring the Limits of Intelligence System: IQ tests

In 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 and AI; rather, it emphasizes the commonality on a more abstract level. For instance, according to [454], 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 space.

The history of analogy can be categorized into three streams of research; see [358] 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 [455]. An example in AI is the *word2vec* model [456, 457], 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 [359]—the Raven's Prograssive Matrices Test (RPM).

RPM has been widely accepted and believed to be highly correlated with real intelligence [458]. Unlike Visual Question Answering (VQA) [459] in computer vision at the periphery of the cognitive ability test circle [458], RPM lies directly at the center of human intelligence, is diagnostic of abstract and structural reasoning ability [460], and characterizes the defining feature of high-level intelligence, *i.e.*, *fluid intelligence* [461]. It has been shown that RPM is harder than existing visual reasoning tests in the following ways [360].

- 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 [458].
- 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) [360] was created to promote further research in this area. The dataset is designed to focus on reasoning and analogy-making instead of visual recognition inherently. 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 38 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

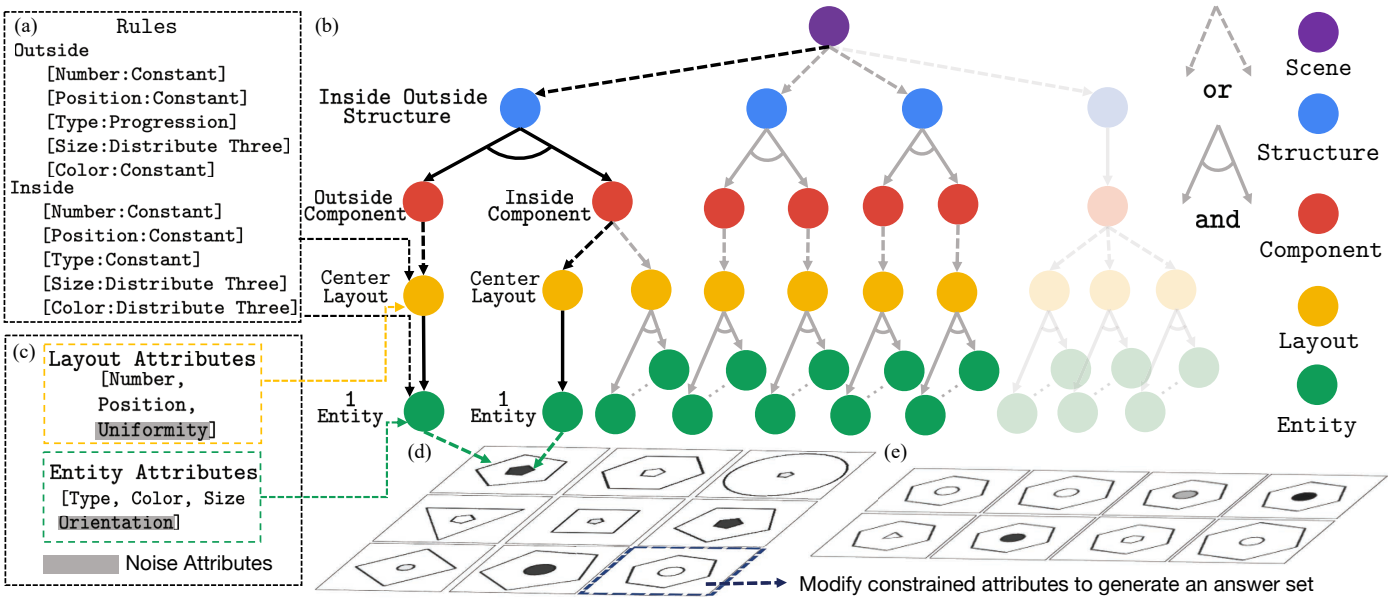


Figure 38: RAVEN creation process proposed in [360]. A graphical illustration of (a) the grammar production rules used in (b) Attributed Stochastic Image Grammar (A-SIG). (c) Note that `Layout` and `Entity` have associated attributes. (d) A sample problem matrix and (e) a sample candidate set.

reasoning. Zhang *et al*. [360] 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 a random guess.

Analogy consists of more than mere spatiotemporal parsing and structural reasoning. For example, *contrast effect* [462] is proven to be one of the key ingredients in relational and analogical reasoning for both human and machine learning [463, 464, 465, 466, 467]. Originated from perceptual learning [468, 469], it is well established in the field of psychology and education [470, 471, 472, 473, 474] that teaching new concepts by comparing with noisy examples is quite effective. Smith and Gentner [475] summarize that comparing cases facilitates transfer learning and problem-solving, as well as the ability to learn relational categories. Gentner [476] in his structure-mapping theory postulate that learners generate a structure alignment between two representation when they compare two cases. A later article [477] 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*. [478] 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*. [479] 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.

9. 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 the following colleagues for helpful discussions on various sections:

- Professor Chenfanfu Jiang at University of Pennsylvania;
- Dr. Behzad Kamgar-Parsi at Office of Naval Research (ONR) and Dr. Bob Madahar at Defence Science and Technology Laboratory (DSTL);
- Luyao Yuan, Zilong Zheng, Xu Xie, Xiaofeng Gao, and Qingyi Zhao at UCLA;
- Dr. Mark Nitzberg, Dr. Mingtian Zhao, and Helen Fu at DMAI, Inc.; and
- Dr. Yibiao Zhao at ISEE, Inc.

[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/`

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.

References

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

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

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

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

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

[6] Sarah H Creem-Regehr, James 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] Mary C Potter, Meaning in visual search, Science 187 (4180) (1975) 965–966.

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

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

[11] Simon Thorpe, Denis Fize, Catherine Marlot, Speed of processing in the human visual system, Nature 381 (6582) (1996) 520.

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

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

[14] Li Fei-Fei, Asha Iyer, Christof Koch, Pietro Perona, What do we perceive in a glance of a real-world scene?, Journal of Vision 7 (1) (2007) 10–10.

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

[16] Aude Oliva, Antonio 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] Arnaud Delorme, Guillaume Richard, Michele 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] Thomas Serre, Aude Oliva, Tomaso Poggio, A feedforward architecture accounts for rapid categorization, Proceedings of the National Academy of Sciences (PNAS) 104 (15) (2007) 6424–6429.

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

[20] Koray Kavukcuoglu, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michaël Mathieu, Yann L Cun, Learning convolutional feature hierarchies for visual recognition, in: Advances in Neural Information Processing Systems (NeurIPS), 2010.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[22] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, James 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] Aude Oliva, Philippe 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] Philippe G Schyns, Diagnostic recognition: task constraints, object information, and their interactions, Cognition 67 (1-2) (1998) 147–179.

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

[26] Siyuan Qi, Siyuan Huang, Ping Wei, Song-Chun Zhu, Predicting human activities using stochastic grammar, in: International Conference on Computer Vision (ICCV), 2017.

[27] Mingtao Pei, Yunde Jia, Song-Chun Zhu, Parsing video events with goal inference and intent prediction, in: International Conference on Computer Vision (ICCV), 2011.

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

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

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

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

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

[33] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, Silvio Savarese, Understanding indoor scenes using 3d geometric phrases, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

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

[35] Xiaobai Liu, Yibiao Zhao, Song-Chun 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] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, Song-Chun Zhu, Holistic 3d scene parsing and reconstruction from a single rgb image, in: European Conference on Computer Vision (ECCV), 2018.

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

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

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

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

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

[42] Patrick Foo, William H Warren, Andrew Duchon, Michael 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] Elizabeth R Chrastil, William H Warren, From cognitive maps to cognitive graphs, PloS one 9 (11) (2014) e112544.

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

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

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

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

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

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

[52] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, Edvard I Moser, Microstructure of a spatial map in the entorhinal cortex, Nature 436 (7052) (2005) 801.

[53] Nathaniel J Killian, Michael J Jutras, Elizabeth A Buffalo, A map of visual space in the primate entorhinal cortex, Nature 491 (7426) (2012) 761.

[54] John O'keefe, Lynn Nadel, The hippocampus as a cognitive map, Oxford: Clarendon Press, 1978.

[55] Joshua Jacobs, Christoph T Weidemann, Jonathan F Miller, Alec Solway, John F Burke, Xue-Xin Wei, Nanthia Suthana, Michael R Sperling, Ashwini D Sharan, Itzhak Fried, et al., Direct recordings of grid-like neuronal activity in human spatial navigation, Nature neuroscience 16 (9) (2013) 1188.

[56] Marianne Fyhn, Torkel Hafting, Menno P Witter, Edvard I Moser, May-Britt Moser, Grid cells in mice, Hippocampus 18 (12) (2008) 1230–1238.

[57] Christian F Doeller, Caswell Barry, Neil Burgess, Evidence for grid cells in a human memory network, Nature 463 (7281) (2010) 657.

[58] Michael M Yartsev, Menno P Witter, Nachum Ulanovsky, Grid cells without theta oscillations in the entorhinal cortex of bats, Nature 479 (7371) (2011) 103.

[59] Ruiqi Gao, Jianwen Xie, Song-Chun Zhu, Ying Nian Wu, Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion, in: International Conference on Learning Rep-

resentations (ICLR), 2019.
[60] Luise Gootjes-Dreesbach, Lyndsey C Pickup, Andrew W Fitzgibbon, Andrew Glennerster, Comparison of view-based and reconstruction-based models of human navigational strategy, Journal of vision 17 (9) (2017) 11–11.
[61] Jenny Vuong, Andrew Fitzgibbon, Andrew Glennerster, Human pointing errors suggest a flattened, task-dependent representation of space, bioRxiv (2018) 390088.
[62] Hoon Choi, Brian J Scholl, Perceiving causality after the fact: Postdiction in the temporal dynamics of causal perception, Perception 35 (3) (2006) 385–399.
[63] Brian J Scholl, Ken Nakayama, Illusory causal crescents: Misperceived spatial relations due to perceived causality, Perception 33 (4) (2004) 455–469.
[64] Brian J Scholl, Tao Gao, Perceiving animacy and intentionality: Visual processing or higher-level judgment, Social perception: Detection and interpretation of animacy, agency, and intention 4629.
[65] Brian J Scholl, Objects and attention: The state of the art, Cognition 80 (1-2) (2001) 1–46.
[66] Ed Vul, George Alvarez, Joshua B Tenenbaum, Michael 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] Peter W Battaglia, Jessica B Hamrick, Joshua B Tenenbaum, Simulation as an engine of physical scene understanding, Proceedings of the National Academy of Sciences (PNAS) 110 (45) (2013) 18327–18332.
[68] Jessica Hamrick, Peter Battaglia, Joshua B Tenenbaum, Internal physics models guide probabilistic judgments about object dynamics, in: Annual Meeting of the Cognitive Science Society (CogSci), 2011.
[69] Dan Xie, Tianmin Shu, Sinisa Todorovic, Song-Chun 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] Tomer Ullman, Andreas Stuhlmüller, Noah Goodman, Joshua B Tenenbaum, Learning physics from dynamical scenes, in: Annual Meeting of the Cognitive Science Society (CogSci), 2014.
[71] Tobias Gerstenberg, Joshua B Tenenbaum, Intuitive theories, in: Oxford handbook of causal reasoning, Oxford University Press New York, NY, 2017, pp. 515–548.
[72] Isaac Newton, John 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] Colin Maclaurin, A Treatise of Fluxions: In Two Books. 1, Vol. 1, Ruddimans, 1742.
[74] Erik T Mueller, Commonsense reasoning: an event calculus based approach, Morgan Kaufmann, 2014.
[75] Erik T Mueller, Daydreaming in humans and machines: a computer model of the stream of thought, Intellect Books, 1990.
[76] Albert Michotte, The perception of causality (TR Miles, Trans.), London, England: Methuen & Co, 1963.
[77] Susan Carey, The origin of concepts, Oxford University Press, 2009.
[78] Ali Farhadi, Ian Endres, Derek Hoiem, David Forsyth, Describing objects by their attributes, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
[79] Devi Parikh, Kristen Grauman, Relative attributes, in: International Conference on Computer Vision (ICCV), 2011.
[80] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, Benjamin Rozenfeld, Learning realistic human actions from movies, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[81] Benjamin Yao, Song-Chun Zhu, Learning deformable action templates from cluttered videos, in: International Conference on Computer Vision (ICCV), 2009.
[82] Benjamin Z Yao, Bruce X Nie, Zicheng Liu, Song-Chun Zhu, Animated pose templates for modeling and detecting human actions, Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 36 (3) (2013) 436–452.
[83] Jiang Wang, Zicheng Liu, Ying Wu, Junsong Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[84] Navneet Dalal, Bill Triggs, Histograms of oriented gradients for human detection, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2005.
[85] Sreemanananth Sadanand, Jason J Corso, Action bank: A high-level representation of activity in video, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[86] Bo Zheng, Yibiao Zhao, C Yu Joey, Katsushi Ikeuchi, Song-Chun Zhu, Detecting potential falling objects by inferring human action and natural disturbance, in: International Conference on Robotics and Automation (ICRA), 2014.
[87] RW Fleming, M Barnett-Cowan, HH Bülthoff, Perceived object stability is affected by the internal representation of gravity, PLoS One 6 (4).
[88] Myrka Zago, Francesco 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] Philip J Kellman, Elizabeth S Spelke, Perception of partly occluded objects in infancy, Cognitive psychology 15 (4) (1983) 483–524.
[90] Renée Baillargeon, Elizabeth S Spelke, Stanley Wasserman, Object permanence in five-month-old infants, Cognition 20 (3) (1985) 191–208.
[91] Scott P Johnson, Richard N Aslin, Perception of object unity in 2-month-old infants, Developmental Psychology 31 (5) (1995) 739.
[92] Amy Needham, Factors affecting infants' use of featural information in object segregation, Current Directions in Psychological Science 6 (2) (1997) 26–33.
[93] Renée Baillargeon, Infants' physical world, Current directions in psychological science 13 (3) (2004) 89–94.
[94] Bo Zheng, Yibiao Zhao, Joey C Yu, Katsushi Ikeuchi, Song-Chun Zhu, Beyond point clouds: Scene understanding by reasoning geometry and physics, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[95] Bo Zheng, Yibiao Zhao, Joey Yu, Katsushi Ikeuchi, Song-Chun Zhu, Scene understanding by reasoning stability and safety, International Journal of Computer Vision (IJCV) (2015) 221–238.
[96] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, Song-Chun Zhu, Human-centric indoor scene synthesis using stochastic grammar, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[97] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, Song-Chun Zhu, Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation, in: Advances in Neural Information Processing Systems (NeurIPS), 2018.
[98] Abhinav Gupta, Scott Satkin, Alexei A Efros, Martial Hebert, From 3d scene geometry to human workspace, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[99] Marco Iacoboni, Istvan Molnar-Szakacs, Vittorio Gallese, Giovanni Buccino, John C Mazziotta, Giacomo Rizzolatti, Grasping the intentions of others with one's own mirror neuron system, PLoS biology 3 (3) (2005) e79.
[100] Gergely Csibra, György Gergely, 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans, Acta psychologica 124 (1) (2007) 60–78.
[101] Chris L Baker, Joshua B Tenenbaum, Rebecca R Saxe, Goal inference as inverse planning, in: Annual Meeting of the Cognitive Science Society (CogSci), 2007.
[102] Chris L Baker, Noah D Goodman, Joshua B Tenenbaum, Theory-based social goal inference, in: Annual Meeting of the Cognitive Science Society (CogSci), 2008.
[103] Minh Hoai, Fernando De la Torre, Max-margin early event detectors, International Journal of Computer Vision (IJCV) 107 (2) (2014) 191–202.
[104] Matthew W Turek, Anthony Hoogs, Roderic Collins, Unsupervised learning of functional categories in video scenes, in: European Conference on Computer Vision (ECCV), 2010.
[105] Helmut Grabner, Juergen Gall, Luc Van Gool, What makes a chair a chair?, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[106] Zhaoyin Jia, Andrew Gallagher, Ashutosh Saxena, Tsuhan Chen, 3d-based reasoning with blocks, support, and stability, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[107] Yun Jiang, Hema Koppula, Ashutosh Saxena, Hallucinated humans as the hidden context for labeling 3d scenes, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[108] Tianmin Shu, Steven M Thurman, Dawn Chen, Song-Chun Zhu, Hongjing Lu, Critical features of joint actions that signal human interaction, in: Annual Meeting of the Cognitive Science Society (CogSci), 2016.
[109] Tianmin Shu, Yujia Peng, Lifeng Fan, Hongjing Lu, Song-Chun Zhu, Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations, Topics in cognitive science 10 (1) (2018) 225–241.
[110] Tianmin Shu, Yujia Peng, Hongjing Lu, Song-Chun Zhu, Partitioning the perception of physical and social events within a unified psychological space, in: Annual Meeting of the Cognitive Science Society (CogSci), 2019.
[111] Chris Baker, Rebecca Saxe, Joshua Tenenbaum, Bayesian theory of mind: Modeling joint belief-desire attribution, in: Annual Meeting of the Cognitive Science Society (CogSci), 2011.
[112] Yibiao Zhao, Steven Holtzen, Tao Gao, Song-Chun Zhu, Represent and infer human theory of mind for human-robot interaction, in: AAAI fall symposium series, 2015.
[113] Noam Nisan, Amir Ronen, Algorithmic mechanism design, Games and Economic behavior 35 (1-2) (2001) 166–196.
[114] Jeremy Bentham, An introduction to the principles of morals, London: Athlone.
[115] Nishant Shukla, Utility learning, non-markovian planning, and task-oriented programming language, Ph.D. thesis, UCLA (2019).
[116] Brian J Scholl, Patrice D Tremoulet, Perceptual causality and animacy, Trends in Cognitive Sciences 4 (8) (2000) 299–309.
[117] Alfred Arthur Robb, Optical geometry of motion: A new view of the theory of relativity, W. Heffer, 1911.
[118] David B Malament, The class of continuous timelike curves determines the topology of spacetime, Journal of mathematical physics 18 (7) (1977) 1399–1404.
[119] Alfred A Robb, Geometry of time and space, Cambridge University Press, 2014.

[120] Roberta Corrigan, Peggy Denton, Causal understanding as a developmental primitive, Developmental review 16 (2) (1996) 162–202.
[121] Peter A White, Causal processing: Origins and development, Psychological bulletin 104 (1) (1988) 36.
[122] Yi-Chia Chen, Brian J Scholl, The perception of history: Seeing causal history in static shapes induces illusory motion perception, Psychological Science 27 (6) (2016) 923–930.
[123] Keith Holyoak, Patricia W. Cheng, Causal learning and inference as a rational process: The new synthesis, Annual Review of Psychology 62 (2011) 135–163.
[124] D. R. Shanks, A. Dickinson, Associative accounts of causality judgment, Psychology of learning and motivation 21 (1988) 229–261.
[125] 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.
[126] Hongjing Lu, Alan L Yuille, Mimi Liljeholm, Patricia W Cheng, Keith J Holyoak, Bayesian generic priors for causal learning, Psychological Review 115 (4) (2008) 955–984.
[127] Mark Edmonds, Siyuan Qi, Yixin Zhu, James Kubricht, Song-Chun Zhu, Hongjing Lu, Decomposing human causal learning: Bottom-up associative learning and top-down schema reasoning, in: Annual Meeting of the Cognitive Science Society (CogSci), 2019.
[128] Michael R Waldmann, Keith J Holyoak, Predictive and diagnostic learning within causal models: asymmetries in cue competition, Journal of Experimental Psychology: General 121 (2) (1992) 222–236.
[129] Mark Edmonds, James Kubricht, Colin Summers, Yixin Zhu, Brandon Rothrock, Song-Chun Zhu, Hongjing Lu, Human causal transfer: Challenges for deep reinforcement learning, in: Annual Meeting of the Cognitive Science Society (CogSci), 2018.
[130] Patricia W Cheng, From covariation to causation: a causal power theory, Psychological Review 104 (2) (1997) 367–405.
[131] Martin Rolfs, Michael Dambacher, Patrick Cavanagh, Visual adaptation of the perception of causality, Current Biology 23 (3) (2013) 250–254.
[132] Celeste McCollough, Color adaptation of edge-detectors in the human visual system, Science 149 (3688) (1965) 1115–1116.
[133] 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.
[134] Tobias Gerstenberg, Matthew F Peterson, Noah D Goodman, David A Lagnado, Joshua B Tenenbaum, Eye-tracking causality, Psychological Science 28 (12) (2017) 1731–1744.
[135] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529.
[136] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, Philipp Moritz, Trust region policy optimization, in: International Conference on Machine Learning (ICML), 2015.
[137] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.
[138] Sergey Levine, Chelsea Finn, Trevor Darrell, Pieter Abbeel, End-to-end training of deep visuomotor policies, The Journal of Machine Learning Research 17 (1) (2016) 1334–1373.
[139] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347.
[140] Chiyuan Zhang, Oriol Vinyals, Remi Munos, Samy Bengio, A study on overfitting in deep reinforcement learning, arXiv preprint arXiv:1804.06893.
[141] Ken Kansky, Tom Silver, David A Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, Dileep George, Schema networks: Zero-shot transfer with a generative causal model of intuitive physics, arXiv preprint arXiv:1706.04317.
[142] Amy Fire, Song-Chun Zhu, Learning perceptual causality from video, ACM Transactions on Intelligent Systems and Technology (TIST) 7 (2) (2016) 23.
[143] Donald B Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies., Journal of educational Psychology 66 (5) (1974) 688.
[144] Guido W Imbens, Donald B Rubin, Causal inference in statistics, social, and biomedical sciences, Cambridge University Press, 2015.
[145] Paul R Rosenbaum, Donald B Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.
[146] J Pearl, Causality: Models, reasoning and inference, Cambridge University Press, 2000.
[147] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, Thomas Richardson, Causation, prediction, and search, MIT press, 2000.
[148] David Maxwell Chickering, Optimal structure identification with greedy search, Journal of machine learning research 3 (Nov) (2002) 507–554.
[149] Jonas Peters, Joris M Mooij, Dominik Janzing, Bernhard Schölkopf, Causal discovery with continuous additive noise models, The Journal of Machine Learning Research 15 (1) (2014) 2009–2053.
[150] Yang-Bo He, Zhi Geng, Active learning of causal networks with intervention experiments and optimal designs, Journal of Machine Learning Research 9 (Nov) (2008) 2523–2547.
[151] Neil R Bramley, Peter Dayan, Thomas L Griffiths, David A Lagnado, Formalizing neurath's ship: Approximate algorithms for online causal learning, Psychological review 124 (3) (2017) 301.
[152] Ronald Aylmer Fisher, The design of experiments, Oliver And Boyd; Edinburgh; London, 1937.
[153] Amy Fire, Song-Chun Zhu, Using causal induction in humans to learn and infer causality from video, in: Annual Meeting of the Cognitive Science Society (CogSci), 2013.
[154] Song-Chun Zhu, Ying Nian Wu, David Mumford, Minimax entropy principle and its application to texture modeling, Neural computation 9 (8) (1997) 1627–1660.
[155] Yuanlu Xu, Lei Qin, Xiaobai Liu, Jianwen Xie, Song-Chun 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.
[156] Caiming Xiong, Nishant Shukla, Wenlong Xiong, Song-Chun Zhu, Robot learning with a spatial, temporal, and causal and-or graph, in: International Conference on Robotics and Automation (ICRA), 2016.
[157] Michael McCloskey, Allyson Washburn, Linda Felch, Intuitive physics: the straight-down belief and its origin, Journal of Experimental Psychology: Learning, Memory, and Cognition 9 (4) (1983) 636.
[158] Michael McCloskey, Alfonso Caramazza, Bert Green, Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects, Science 210 (4474) (1980) 1139–1141.
[159] Andrea A DiSessa, Unlearning aristotelian physics: A study of knowledge-based learning, Cognitive science 6 (1) (1982) 37–75.
[160] Mary Kister Kaiser, John Jonides, Joanne Alexander, Intuitive reasoning about abstract and familiar physics problems, Memory & Cognition 14 (4) (1986) 308–312.
[161] Kevin A Smith, Peter Battaglia, Edward Vul, Consistent physics underlying ballistic motion prediction, in: Annual Meeting of the Cognitive Science Society (CogSci), 2013.
[162] Mary K Kaiser, Dennis R Proffitt, Susan M Whelan, Heiko Hecht, Influence of animation on dynamical judgments, Journal of experimental Psychology: Human Perception and performance 18 (3) (1992) 669.
[163] Mary K Kaiser, Dennis R Proffitt, Kenneth 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.
[164] In-Kyeong Kim, Elizabeth S Spelke, Perception and understanding of effects of gravity and inertia on object motion, Developmental Science 2 (3) (1999) 339–362.
[165] Jean Piaget, Margaret Cook, The origins of intelligence in children, International Universities Press New York, 1952.
[166] Jean Piaget, Margaret Trans Cook, The construction of reality in the child., Basic Books, 1954.
[167] Susan J Hespos, Renée Baillargeon, Décalage in infants' knowledge about occlusion and containment events: Converging evidence from action tasks, Cognition 99 (2) (2006) B31–B41.
[168] Susan J Hespos, Renée Baillargeon, Young infants' actions reveal their developing knowledge of support variables: Converging evidence for violation-of-expectation findings, Cognition 107 (1) (2008) 304–316.
[169] T GR Bower, Development in infancy, WH Freeman, 1974.
[170] Alan M Leslie, Stephanie Keeble, Do six-month-old infants perceive causality?, Cognition 25 (3) (1987) 265–288.
[171] Yuyan Luo, Renée Baillargeon, Laura Brueckner, Yuko Munakata, Reasoning about a hidden object after a delay: Evidence for robust representations in 5-month-old infants, Cognition 88 (3) (2003) B23–B32.
[172] Renée Baillargeon, Jie Li, Weiting Ng, Sylvia Yuan, An account of infants' physical reasoning, in: Learning and the Infant Mind, Oxford University Press, 2008, pp. 66–116.
[173] Renée Baillargeon, The acquisition of physical knowledge in infancy: A summary in eight lessons, Blackwell handbook of childhood cognitive development 1 (46-83) (2002) 1.
[174] Yixin Chen, Siyuan Huang, Tao Yuan, Yixin Zhu, Siyuan Qi, Song-Chun Zhu, Holistic++ scene understanding with human-object interaction and physical commonsense, in: International Conference on Computer Vision (ICCV), 2019.
[175] Peter Achinstein, The nature of explanation, Oxford University Press on Demand, 1983.
[176] Jason Fischer, John G Mikhael, Joshua B Tenenbaum, Nancy Kanwisher, Functional neuroanatomy of intuitive physical inference, Proceedings of the National Academy of Sciences (PNAS) 113 (34) (2016) E5072–E5081.
[177] Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, Joshua B Tenenbaum, Mind games: Game engines as an architecture for intuitive physics, Trends in Cognitive Sciences 21 (9) (2017) 649–665.
[178] Christopher Bates, Peter Battaglia, Ilker Yildirim, Joshua B Tenenbaum, Humans predict liquid dynamics using probabilistic simulation, in: Annual Meeting of the Cognitive Science Society (CogSci), 2015.
[179] James Kubricht, Chenfanfu Jiang, Yixin Zhu, Song-Chun Zhu, Demetri

Terzopoulos, Hongjing Lu, Probabilistic simulation predicts human performance on viscous fluid-pouring problem, in: Annual Meeting of the Cognitive Science Society (CogSci), 2016.
[180] James Kubricht, Yixin Zhu, Chenfanfu Jiang, Demetri Terzopoulos, Song-Chun Zhu, Hongjing Lu, Consistent probabilistic simulation underlying human judgment in substance dynamics, in: Annual Meeting of the Cognitive Science Society (CogSci), 2017.
[181] James R Kubricht, Keith J Holyoak, Hongjing Lu, Intuitive physics: Current research and controversies, Trends in Cognitive Sciences 21 (10) (2017) 749–759.
[182] David Mumford, Agnès Desolneux, Pattern theory: the stochastic analysis of real-world signals, AK Peters/CRC Press, 2010.
[183] David Mumford, Pattern theory: a unifying perspective, in: First European congress of mathematics, Springer, 1994.
[184] Bela Julesz, Visual pattern discrimination, IRE transactions on Information Theory 8 (2) (1962) 84–92.
[185] Song-Chun Zhu, Yingnian Wu, David 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.
[186] Bela Julesz, Textons, the elements of texture perception, and their interactions, Nature 290 (5802) (1981) 91.
[187] Song-Chun Zhu, Cheng-En Guo, Yizhou Wang, Zijian Xu, What are textons?, International Journal of Computer Vision (IJCV) 62 (1-2) (2005) 121–143.
[188] Cheng-en Guo, Song-Chun Zhu, Ying Nian Wu, Towards a mathematical theory of primal sketch and sketchability, in: International Conference on Computer Vision (ICCV), 2003.
[189] Cheng-en Guo, Song-Chun Zhu, Ying Nian Wu, Primal sketch: Integrating structure and texture, Computer Vision and Image Understanding (CVIU) 106 (1) (2007) 5–19.
[190] Mark Nitzberg, David Mumford, The 2.1-d sketch, in: ICCV, 1990.
[191] John YA Wang, Edward H Adelson, Layered representation for motion analysis, in: Conference on Computer Vision and Pattern Recognition (CVPR), 1993.
[192] John YA Wang, Edward H Adelson, Representing moving images with layers, Transactions on Image Processing (TIP) 3 (5) (1994) 625–638.
[193] David Marr, Herbert Keith 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.
[194] I Binford, Visual perception by computer, in: IEEE Conference of Systems and Control, 1971.
[195] Rodney A Brooks, Symbolic reasoning among 3-d models and 2-d images, Artificial Intelligence 17 (1-3) (1981) 285–348.
[196] Takeo Kanade, Recovery of the three-dimensional shape of an object from a single view, Artificial intelligence 17 (1-3) (1981) 409–460.
[197] Donald Broadbent, A question of levels: Comment on McClelland and Rumelhart, American Psychological Association, 1985.
[198] David Lowe, Perceptual organization and visual recognition, Vol. 5, Springer Science & Business Media, 2012.
[199] Alex P Pentland, Perceptual organization and the representation of natural form, in: Readings in Computer Vision, Elsevier, 1987, pp. 680–699.
[200] Kurt Koffka, Principles of Gestalt psychology, Routledge, 2013.
[201] David Waltz, Understanding line drawings of scenes with shadows, in: The psychology of computer vision, 1975.
[202] Harry G Barrow, Jay M Tenenbaum, Interpreting line drawings as three-dimensional surfaces, Artificial Intelligence 17 (1-3) (1981) 75–116.
[203] David G Lowe, Three-dimensional object recognition from single two-dimensional images, Artificial Intelligence 31 (3) (1987) 355–395.
[204] David G Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.
[205] Robert L Solso, M Kimberly MacLin, Otto H MacLin, Cognitive psychology, Pearson Education New Zealand, 2005.
[206] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, Josh Tenenbaum, Galileo: Perceiving physical object properties by integrating a physics engine with deep learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2015.
[207] Peter Dayan, Geoffrey E Hinton, Radford M Neal, Richard S Zemel, The helmholtz machine, Neural computation 7 (5) (1995) 889–904.
[208] Lawrence G Roberts, Machine perception of three-dimensional solids, Ph.D. thesis, Massachusetts Institute of Technology (1963).
[209] Irving Biederman, Robert J Mezzanotte, Jan C Rabinowitz, Scene perception: Detecting and judging objects undergoing relational violations, Cognitive psychology (1982) 143–177.
[210] Manuel Blum, Arnold Griffith, Bernard Neumann, A stability test for configurations of blocks, Tech. rep., Massachusetts Institute of Technology (1970).
[211] Matthew Brand, Paul Cooper, Lawrence Birnbaum, Seeing physics, or: Physics is for prediction, in: Proceedings of the Workshop on Physics-based Modeling in Computer Vision, 1995.
[212] Abhinav Gupta, Alexei A Efros, Martial Hebert, Blocks world revisited: Image understanding using qualitative geometry and mechanics, in: European Conference on Computer Vision (ECCV), 2010.
[213] Varsha Hedau, Derek Hoiem, David Forsyth, Recovering the spatial layout of cluttered rooms, in: International Conference on Computer Vision (ICCV), 2009.
[214] David C Lee, Martial Hebert, Takeo Kanade, Geometric reasoning for single image structure recovery, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
[215] Varsha Hedau, Derek Hoiem, David Forsyth, Recovering free space of indoor scenes from a single image, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[216] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, Rob Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision (ECCV), 2012.
[217] Alexander G Schwing, Tamir Hazan, Marc Pollefeys, Raquel Urtasun, Efficient structured prediction for 3d indoor scene understanding, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[218] Ruiqi Guo, Derek Hoiem, Support surface prediction in indoor scenes, in: International Conference on Computer Vision (ICCV), 2013.
[219] Tianjia Shao, Aron Monszpart, Youyi Zheng, Bongjin Koo, Weiwei Xu, Kun Zhou, Niloy J Mitra, Imagining the unseen: Stability-based cuboid arrangements for scene understanding, ACM Transactions on Graphics (TOG) 33 (6).
[220] Yilun Du, Zhijian Liu, Hector Basevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, Jiajun Wu, Learning to exploit stability for 3d scene parsing, in: Advances in Neural Information Processing Systems (NeurIPS), 2018.
[221] Yixin Zhu, Yibiao Zhao, Song-Chun Zhu, Understanding tools: Task-oriented object modeling, learning and recognition, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[222] Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, William T Freeman, Physics 101: Learning physical object properties from unlabeled videos, in: British Machine Vision Conference (BMVC), 2016.
[223] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, Song-Chun Zhu, Inferring forces and learning human utilities from videos, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[224] Marcus A Brubaker, David J Fleet, The kneed walker for human pose tracking, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[225] Marcus A Brubaker, Leonid Sigal, David J Fleet, Estimating contact dynamics, in: International Conference on Computer Vision (ICCV), 2009.
[226] Marcus A Brubaker, David J Fleet, Aaron Hertzmann, Physics-based person tracking using the anthropomorphic walker, International Journal of Computer Vision (IJCV) 87 (1-2) (2010) 140.
[227] Tu-Hoa Pham, Abderrahmane Kheddar, Ammar Qammaz, Antonis 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.
[228] Yangang Wang, Jianyuan Min, Jianjie Zhang, Yebin Liu, Feng Xu, Qionghai Dai, Jinxiang Chai, Video-based hand manipulation capture through composite motion control, ACM Transactions on Graphics (TOG) 32 (4) (2013) 43.
[229] Wenping Zhao, Jianjie Zhang, Jianyuan Min, Jinxiang Chai, Robust realtime physics-based motion control for human grasping, ACM Transactions on Graphics (TOG) 32 (6) (2013) 207.
[230] James J Gibson, The perception of the visual world, Houghton Mifflin, 1950.
[231] James Jerome Gibson, The senses considered as perceptual systems, Houghton Mifflin, 1966.
[232] Katherine Nelson, Concept, word, and sentence: interrelations in acquisition and development, Psychological review 81 (4) (1974) 267.
[233] James J Gibson, The theory of affordances, Hilldale, USA.
[234] Mohammed Hassanin, Salman Khan, Murat Tahtali, Visual affordance and function understanding: A survey, arXiv preprint arXiv:1807.06775.
[235] Huaqing Min, Chang'an Yi, Ronghua Luo, Jinhui Zhu, Sheng Bi, Affordance research in developmental robotics: A survey, IEEE Transactions on Cognitive and Developmental Systems 8 (4) (2016) 237–255.
[236] Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, Data-driven grasp synthesis—a survey, IEEE Transactions on Robotics 30 (2) (2013) 289–309.
[237] Natsuki Yamanobe, Weiwei Wan, Ixchel G Ramirez-Alpizar, Damien Petit, Tokuo Tsuji, Shuichi Akizuki, Manabu Hashimoto, Kazuyuki Nagata, Kensuke Harada, A brief review of affordance in robotic manipulation research, Advanced Robotics 31 (19-20) (2017) 1086–1101.
[238] Wolfgang Kohler, The mentality of apes, New York: Liverright, 1925.
[239] William Homan Thorpe, Learning and instinct in animals, Harvard University Press, 1956.
[240] Kenneth Page Oakley, Man the tool-maker, University of Chicago Press, 1968.
[241] Jane Goodall, The Chimpanzees of Gombe: Patterns of Behavior, Bellknap Press of the Harvard University Press, 1986.
[242] Andrew Whiten, Jane Goodall, William C McGrew, Toshisada Nishida, Vernon Reynolds, Yukimaru Sugiyama, Caroline EG Tutin, Richard W Wrangham, Christophe Boesch, Cultures in chimpanzees, Nature 399 (6737) (1999) 682.
[243] Richard W Byrne, Andrew Whiten, Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans, Clarendon Press/Oxford University Press, 1988.
[244] Gloria Sabbatini, Héctor Marín Manrique, Cinzia Trapanese, Aurora

De Bortoli Vizioli, Josep Call, Elisabetta Visalberghi, Sequential use of rigid and pliable tools in tufted capuchin monkeys (sapajus spp.), Animal Behaviour 87 (2014) 213–220.
[245] Gavin R Hunt, Manufacture and use of hook-tools by new caledonian crows, Nature 379 (6562) (1996) 249.
[246] Alex AS Weir, Jackie Chappell, Alex Kacelnik, Shaping of hooks in new caledonian crows, Science 297 (5583) (2002) 981–981.
[247] Dakota E McCoy, Martina Schiestl, Patrick Neilands, Rebecca Hassall, Russell D Gray, Alex H Taylor, New caledonian crows behave optimistically after using tools, Current Biology 29 (16) (2019) 2737–2742.
[248] Benjamin B Beck, Animal tool behavior: The use and manufacture of tools by animals, Garland STPM Press New York, 1980.
[249] Christopher D Bird, Nathan 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.
[250] Peter Freeman, Allen Newell, A model for functional reasoning in design, in: International Joint Conference on Artificial Intelligence (IJCAI), 1971.
[251] Patrick H Winston, Learning structural descriptions from examples, Tech. rep., Massachusetts Institute of Technology (1970).
[252] Patrick H Winston, Thomas O Binford, Boris Katz, Michael Lowry, Learning physical descriptions from functional definitions, examples, and precedents, in: AAAI Conference on Artificial Intelligence (AAAI), 1983.
[253] Michael Brady, Philip E. Agre, The mechanic's mate, in: Advances in Artificial Intelligence, Proceedings of the Sixth European Conference on Artificial Intelligence (ECAI), 1984.
[254] Jonathan H Connell, Michael Brady, Generating and generalizing models of visual objects, Artificial Intelligence 31 (2) (1987) 159–183.
[255] Seng-Beng Ho, Representing and using functional definitions for visual recognition, Ph.D. thesis, The University of Wisconsin-Madison (1987).
[256] M DiManzo, Emanuele Trucco, Fausto Giunchiglia, F Ricci, Fur: Understanding functional reasoning, International Journal of Intelligent Systems 4 (4) (1989) 431–457.
[257] Marvin Minsky, Society of mind, Simon and Schuster, 1988.
[258] Louise Stark, Kevin 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.
[259] Zhijian Liu, William T Freeman, Joshua B Tenenbaum, Jiajun Wu, Physical primitive decomposition, in: European Conference on Computer Vision (ECCV), 2018.
[260] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, Joseph J Lim, Demo2vec: Reasoning object affordances from online videos, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[261] Christopher Baber, Cognition and tool use: Forms of engagement in human and animal use of tools, CRC Press, 2003.
[262] Bärbel Inhelder, Jean Piaget, The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures, Vol. 22, Psychology Press, 1958.
[263] Brent Strickland, Brian J Scholl, Visual perception involves event-type representations: The case of containment versus occlusion, Journal of Experimental Psychology: General 144 (3) (2015) 570.
[264] Marianella Casasola, Leslie B Cohen, Infant categorization of containment, support and tight-fit spatial relationships, Developmental Science 5 (2) (2002) 247–264.
[265] Susan J Hespos, Renée Baillargeon, Reasoning about containment events in very young infants, Cognition 78 (3) (2001) 207–245.
[266] Su-hua Wang, Renée Baillargeon, Sarah Paterson, Detecting continuity violations in infancy: A new account and new evidence from covering and tube events, Cognition 95 (2) (2005) 129–173.
[267] Susan J Hespos, Elizabeth 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.
[268] Ernest Davis, Gary Marcus, Noah Frazier-Logue, Commonsense reasoning about containers using radically incomplete information, Artificial intelligence 248 (2017) 46–84.
[269] Ernest Davis, How does a box work? a study in the qualitative dynamics of solid objects, Artificial Intelligence 175 (1) (2011) 299–345.
[270] Ernest Davis, Pouring liquids: A study in commonsense physical reasoning, Artificial Intelligence 172 (12-13) (2008) 1540–1578.
[271] Anthony G Cohn, Qualitative spatial representation and reasoning techniques, in: Annual Conference on Artificial Intelligence, Springer, 1997.
[272] Anthony G. Cohn, Shyamanta M. Hazarika, Qualitative spatial representation and reasoning: An overview, Fundamenta informaticae 46 (1-2) (2001) 1–29.
[273] Wei Liang, Yibiao Zhao, Yixin Zhu, Song-Chun Zhu, Evaluating human cognition of containing relations with physical simulation, in: Annual Meeting of the Cognitive Science Society (CogSci), 2015.
[274] Lap-Fai Yu, Noah Duncan, Sai-Kit Yeung, Fill and transfer: A simple physics-based approach for containability reasoning, in: International Conference on Computer Vision (ICCV), 2015.
[275] Roozbeh Mottaghi, Connor Schenck, Dieter Fox, Ali Farhadi, See the glass half full: Reasoning about liquid containers, their volume and content, in: International Conference on Computer Vision (ICCV), 2017.
[276] Wei Liang, Yibiao Zhao, Yixin Zhu, Song-Chun Zhu, What is where: Inferring containment relations from videos, in: International Joint Conference on Artificial Intelligence (IJCAI), 2016.
[277] Wei Liang, Yixin Zhu, Song-Chun Zhu, Tracking occluded objects and recovering incomplete trajectories by reasoning about containment relations and human actions, in: AAAI Conference on Artificial Intelligence (AAAI), 2018.
[278] Yun Jiang, Marcus Lim, Ashutosh Saxena, Learning object arrangements in 3d scenes using human context, in: International Conference on Machine Learning (ICML), 2012.
[279] Chenfanfu Jiang, Siyuan Qi, Yixin Zhu, Siyuan Huang, Jenny Lin, Lap-Fai Yu, Demetri Terzopoulos, Song-Chun 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.
[280] Kerstin Dautenhahn, Chrystopher L Nehaniv, Imitation in Animals and Artifacts, MIT Press Cambridge, MA, 2002.
[281] Brenna D Argall, Sonia Chernova, Manuela Veloso, Brett Browning, A survey of robot learning from demonstration, Robotics and Autonomous Systems 57 (5) (2009) 469–483.
[282] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al., An algorithmic perspective on imitation learning, Foundations and Trends® in Robotics 7 (1-2) (2018) 1–179.
[283] Ye Gu, Weihua Sheng, Meiqin Liu, Yongsheng Ou, Fine manipulative action recognition through sensor fusion, in: International Conference on Intelligent Robots and Systems (IROS), 2015.
[284] Frank L Hammond, Yiğit Mengüç, Robert 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.
[285] Hangxin Liu, Xu Xie, Matt Millar, Mark Edmonds, Feng Gao, Yixin Zhu, Veronica J Santos, Brandon Rothrock, Song-Chun 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.
[286] Mark Edmonds, Feng Gao, Xu Xie, Hangxin Liu, Siyuan Qi, Yixin Zhu, Brandon Rothrock, Song-Chun 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.
[287] Hangxin Liu, Yaofang Zhang, Wenwen Si, Xu Xie, Yixin Zhu, Song-Chun Zhu, Interactive robot knowledge patching using augmented reality, in: International Conference on Robotics and Automation (ICRA), 2018.
[288] Hangxin Liu, Chi Zhang, Yixin Zhu, Chenfanfu Jiang, Song-Chun Zhu, Mirroring without overimitation: Learning functionally equivalent manipulation actions, in: AAAI Conference on Artificial Intelligence (AAAI), 2019.
[289] Daniel Clement Dennett, The intentional stance, MIT press, 1989.
[290] Fritz Heider, The psychology of interpersonal relations, Psychology Press, 2013.
[291] György Gergely, Zoltán Nádasdy, Gergely Csibra, Szilvia Bíró, Taking the intentional stance at 12 months of age, Cognition 56 (2) (1995) 165–193.
[292] David Premack, Guy Woodruff, Does the chimpanzee have a theory of mind?, Behavioral and brain sciences 1 (4) (1978) 515–526.
[293] Dare A Baldwin, Jodie A Baird, Discerning intentions in dynamic human action, Trends in Cognitive Sciences 5 (4) (2001) 171–178.
[294] Amanda L Woodward, Infants selectively encode the goal object of an actor's reach, Cognition 69 (1) (1998) 1–34.
[295] Andrew N Meltzoff, Rechele Brooks, Like me" as a building block for understanding other minds: Bodily acts, attention, and intention, Intentions and intentionality: Foundations of social cognition 171191.
[296] Dare A Baldwin, Jodie A Baird, Megan M Saylor, M Angela Clark, Infants parse dynamic action, Child development 72 (3) (2001) 708–717.
[297] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, Henrike Moll, Understanding and sharing intentions: The origins of cultural cognition, Behavioral and brain sciences 28 (5) (2005) 675–691.
[298] Szilvia Biro, Bernhard Hommel, Becoming an intentional agent: introduction to the special issue., Acta psychologica 124 (1) (2007) 1–7.
[299] György Gergely, Harold Bekkering, Ildikó Király, Developmental psychology: Rational imitation in preverbal infants, Nature 415 (6873) (2002) 755.
[300] Amanda L Woodward, Jessica A Sommerville, Sarah Gerson, Annette ME Henderson, Jennifer Buresh, The emergence of intention attribution in infancy, Psychology of learning and motivation 51 (2009) 187–222.
[301] M Tomasello, Developing theories of intention (1999).
[302] Paul Bloom, Intention, history, and artifact concepts, Cognition 60 (1) (1996) 1–29.
[303] Fritz Heider, Marianne Simmel, An experimental study of apparent behavior, The American journal of psychology 57 (2) (1944) 243–259.
[304] Diane S Berry, Stephen J Misovich, Methodological approaches to the study of social event perception, Personality and Social Psychology Bulletin 20 (2) (1994) 139–152.
[305] John N Bassili, Temporal and spatial contingencies in the perception of social events, Journal of Personality and Social Psychology 33 (6) (1976) 680.

[306] Winand H Dittrich, Stephen EG Lea, Visual perception of intentional motion, Perception 23 (3) (1994) 253–268.
[307] Daniel C Dennett, Précis of the intentional stance, Behavioral and brain sciences 11 (3) (1988) 495–505.
[308] Shari Liu, Neon B Brooks, Elizabeth S Spelke, Origins of the concepts cause, cost, and goal in prereaching infants, Proceedings of the National Academy of Sciences (PNAS) (2019) 201904410.
[309] Tao Gao, George E Newman, Brian J Scholl, The psychophysics of chasing: A case study in the perception of animacy, Cognitive psychology 59 (2) (2009) 154–179.
[310] Steven Holtzen, Yibiao Zhao, Tao Gao, Joshua B Tenenbaum, Song-Chun Zhu, Inferring human intent from video by sampling hierarchical plans, in: International Conference on Intelligent Robots and Systems (IROS), 2016.
[311] Shari Liu, Elizabeth S Spelke, Six-month-old infants expect agents to minimize the cost of their actions, Cognition 160 (2017) 35–42.
[312] György Gergely, Gergely Csibra, Teleological reasoning in infancy: The naıve theory of rational action, Trends in Cognitive Sciences 7 (7) (2003) 287–292.
[313] Chris L Baker, Rebecca Saxe, Joshua B Tenenbaum, Action understanding as inverse planning, Cognition 113 (3) (2009) 329–349.
[314] Luís Moniz Pereira, et al., Intention recognition via causal bayes networks plus plan generation, in: Portuguese Conference on Artificial Intelligence, Springer, 2009.
[315] Sahil Narang, Andrew Best, Dinesh 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.
[316] Ryo Nakahashi, Chris L Baker, Joshua B Tenenbaum, Modeling human understanding of complex intentional action with a bayesian nonparametric subgoal model, in: AAAI Conference on Artificial Intelligence (AAAI), 2016.
[317] Yu Kong, Yun Fu, Human action recognition and prediction: A survey, arXiv preprint arXiv:1806.11230.
[318] Sarah-Jayne Blakemore, Jean Decety, From the perception of action to the understanding of intention, Nature reviews neuroscience 2 (8) (2001) 561.
[319] Birgit Elsner, Bernhard Hommel, Effect anticipation and action control., Journal of experimental psychology: human perception and performance 27 (1) (2001) 229.
[320] Birgit Elsner, Infants' imitation of goal-directed actions: The role of movements and action effects, Acta psychologica 124 (1) (2007) 44–59.
[321] Giacomo Rizzolatti, Laila Craighero, The mirror-neuron system, Annual Review of Neuroscience 27 (2004) 169–192.
[322] Jonas T Kaplan, Marco Iacoboni, Getting a grip on other minds: Mirror neurons, intention understanding, and cognitive empathy, Social neuroscience 1 (3-4) (2006) 175–183.
[323] Vincent M Reid, Gergely Csibra, Jay Belsky, Mark H Johnson, Neural correlates of the perception of goal-directed action in infants, Acta psychologica 124 (1) (2007) 129–138.
[324] Gergely Csibra, György Gergely, The teleological origins of mentalistic action explanations: A developmental hypothesis, Developmental Science 1 (2) (1998) 255–259.
[325] György Gergely, The development of understanding self and agency, Blackwell handbook of childhood cognitive development (2002) 26–46.
[326] Chris L Kleinke, Gaze and eye contact: a research review, Psychological bulletin 100 (1) (1986) 78.
[327] Nathan J Emery, The eyes have it: the neuroethology, function and evolution of social gaze, Neuroscience & Biobehavioral Reviews 24 (6) (2000) 581–604.
[328] Judee K Burgoon, Laura K Guerrero, Kory Floyd, Nonverbal communication, Routledge, 2016.
[329] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, Song-Chun 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.
[330] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, Song-Chun Zhu, Understanding human gaze communication by spatio-temporal graph reasoning, in: International Conference on Computer Vision (ICCV), 2019.
[331] Alicia P Melis, Michael Tomasello, Chimpanzees (pan troglodytes) coordinate by communicating in a collaborative problem-solving task, Proceedings of the Royal Society B 286 (1901) (2019) 20190408.
[332] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, Song-Chun Zhu, Inferring shared attention in social scene videos, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[333] Susanne Trick, Dorothea Koert, Jan Peters, Constantin Rothkopf, Multimodal uncertainty reduction for intention recognition in human-robot interaction, arXiv preprint arXiv:1907.02426.
[334] Tianmin Shu, Michael S Ryoo, Song-Chun Zhu, Learning social affordance for human-robot interaction, in: International Joint Conference on Artificial Intelligence (IJCAI), 2016.
[335] Tianmin Shu, Xiaofeng Gao, Michael S Ryoo, Song-Chun Zhu, Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions, in: International Conference on Robotics and Automation (ICRA), 2017.
[336] Stuart J Russell, Peter Norvig, Artificial intelligence: a modern approach, Malaysia; Pearson Education Limited,, 2016.
[337] Francis Hutcheson, An Inquiry into the Original of our Ideas of Beauty and Virtue: in two treatises, J. Darby...[and 8 others], 1726.
[338] John Stuart Mill, Utilitarianism, Longmans, Green and Company, 1863.
[339] Xu Xie, Hangxin Liu, Zhenliang Zhang, Yuxing Qiu, Feng Gao, Siyuan Qi, Yixin Zhu, Song-Chun Zhu, Vrgym: A virtual testbed for physical and interactive ai, in: Proceedings of the ACM TURC, 2019.
[340] Nishant Shukla, Yunzhong He, Frank Chen, Song-Chun Zhu, Learning human utility from video demonstrations for deductive planning in robotics, in: Conference on Robot Learning, 2017.
[341] H Paul Grice, Peter Cole, Jerry Morgan, et al., Logic and conversation, 1975 (1975) 41–58.
[342] Noah D Goodman, Michael C Frank, Pragmatic language interpretation as probabilistic inference, Trends in Cognitive Sciences 20 (11) (2016) 818–829.
[343] David Lewis, Convention: A philosophical study, John Wiley & Sons, 2008.
[344] Dan Sperber, Deirdre Wilson, Relevance: Communication and cognition, Vol. 142, Harvard University Press Cambridge, MA, 1986.
[345] Ludwig Wittgenstein, Philosophical Investigations, Macmillan, 1953.
[346] Herbert H Clark, Using language, Cambridge university press, 1996.
[347] Ciyang Qing, Michael Franke, Variations on a bayesian theme: Comparing bayesian models of referential reasoning, in: Bayesian natural language semantics and pragmatics, Springer, 2015, pp. 201–220.
[348] Noah D Goodman, Andreas Stuhlmüller, Knowledge and implicature: Modeling language understanding as social cognition, Topics in cognitive science 5 (1) (2013) 173–184.
[349] Robert Dale, Ehud Reiter, Computational interpretations of the gricean maxims in the generation of referring expressions, Cognitive science 19 (2) (1995) 233–263.
[350] Anton Benz, Gerhard Jäger, Robert Van Rooij, An introduction to game theory for linguists, in: Game theory and pragmatics, Springer, 2006, pp. 1–82.
[351] Gerhard Jäger, Applications of game theory in linguistics, Language and Linguistics compass 2 (3) (2008) 406–421.
[352] Michael C Frank, Noah D Goodman, Predicting pragmatic reasoning in language games, Science 336 (6084) (2012) 998–998.
[353] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, Igor Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, in: Advances in Neural Information Processing Systems (NeurIPS), 2017.
[354] Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, Shimon Whiteson, Learning to communicate with deep multi-agent reinforcement learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2016.
[355] Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, Shimon Whiteson, Stabilising experience replay for deep multi-agent reinforcement learning, in: International Conference on Machine Learning (ICML), 2017.
[356] Prashant Doshi, Piotr J Gmytrasiewicz, Monte carlo sampling methods for approximating interactive pomdps, Journal of Artificial Intelligence Research 34 (2009) 297–337.
[357] Michael Kinney, Costas Tsatsoulis, Learning communication strategies in multiagent systems, Applied intelligence 9 (1) (1998) 71–91.
[358] Keith J Holyoak, Analogy and relational reasoning, in: The Oxford Handbook of Thinking and Reasoning, Oxford University Press, 2012, pp. 234–259.
[359] J. C. et al. Raven, Raven's progressive matrices, Western Psychological Services.
[360] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, Song-Chun Zhu, Raven: A dataset for relational and analogical visual reasoning, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[361] Shane Legg, Marcus Hutter, Universal intelligence: A definition of machine intelligence, Minds and machines 17 (4) (2007) 391–444.
[362] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, Hao 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.
[363] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al., Shapenet: An information-rich 3d model repository, arXiv preprint arXiv:1512.03012.
[364] Tian Feng, Lap-Fai Yu, Sai-Kit Yeung, KangKang Yin, Kun Zhou, Crowd-driven mid-scale layout design., ACM Transactions on Graphics (TOG) 35 (4) (2016) 132–1.
[365] Manolis Savva, Angel X Chang, Alexey Dosovitskiy, Thomas Funkhouser, Vladlen Koltun, Minos: Multimodal indoor simulator for navigation in complex environments, arXiv preprint arXiv:1712.03931.
[366] Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, Florian Strub, Jean Rouat, Hugo Larochelle, Aaron Courville, Home: A household multimodal environment, arXiv preprint arXiv:1711.11017.
[367] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, Silvio Savarese, Gibson env: Real-world perception for embodied agents, in: Conference on Computer Vision and Pattern Recognition (CVPR),

2018.
[368] Yi Wu, Yuxin Wu, Georgia Gkioxari, Yuandong Tian, Building generalizable agents with a realistic and rich 3d environment, arXiv preprint arXiv:1801.02209.
[369] Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, Ali Farhadi, Ai2-thor: An interactive 3d environment for visual ai, arXiv preprint arXiv:1712.05474.
[370] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, Antonio Torralba, Virtualhome: Simulating household activities via programs, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[371] Xiaofeng Gao, Ran Gong, Tianmin Shu, Xu Xie, Shu Wang, Song-Chun Zhu, Vrkitchen: an interactive 3d virtual environment for task-oriented learning, arXiv preprint arXiv:1903.05757.
[372] Shital Shah, Debadeepta Dey, Chris Lovett, Ashish Kapoor, Airsim: High-fidelity visual and physical simulation for autonomous vehicles, in: Field and service robotics, Springer, 2018.
[373] Ming Gao, Xinlei Wang, Kui Wu, Andre Pradhana, Eftychios Sifakis, Cem Yuksel, Chenfanfu Jiang, Gpu optimization of material point methods, ACM Transactions on Graphics (TOG) 37 (6).
[374] Demetri Terzopoulos, John Platt, Alan Barr, Kurt Fleischer, Elastically deformable models, ACM Transactions on Graphics (TOG) 21 (4) (1987) 205–214.
[375] Demetri Terzopoulos, Kurt Fleischer, Modeling inelastic deformation: viscolelasticity, plasticity, fracture, ACM Transactions on Graphics (TOG) 22 (4) (1988) 269–278.
[376] Nick Foster, Dimitri Metaxas, Realistic animation of liquids, Graphical models and image processing 58 (5) (1996) 471–483.
[377] Jos Stam, Stable fluids, in: ACM Transactions on Graphics (TOG), Vol. 99, 1999.
[378] Robert Bridson, Fluid simulation for computer graphics, CRC Press, 2015.
[379] Javier Bonet, Richard D Wood, Nonlinear continuum mechanics for finite element analysis, Cambridge university press, 1997.
[380] 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.
[381] Jan Hegemann, Chenfanfu Jiang, Craig Schroeder, Joseph M Teran, A level set method for ductile fracture, in: ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA), 2013.
[382] Theodore F Gast, Craig Schroeder, Alexey Stomakhin, Chenfanfu Jiang, Joseph M Teran, Optimization integrator for large time steps, IEEE Transactions on Visualization and Computer Graph (TVCG) 21 (10) (2015) 1103–1115.
[383] Minchen Li, Ming Gao, Timothy Langlois, Chenfanfu Jiang, Danny M Kaufman, Decomposed optimization time integrator for large-step elastodynamics, ACM Transactions on Graphics (TOG) 38 (4) (2019) 70.
[384] Yuting Wang, Chenfanfu Jiang, Craig Schroeder, Joseph Teran, An adaptive virtual node algorithm with robust mesh cutting, in: ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA), 2014.
[385] J. J. Monaghan, Smoothed particle hydrodynamics, Annual review of astronomy and astrophysics 30 (1) (1992) 543–574.
[386] 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.
[387] S. Li, W. K. Liu, Meshfree and particle methods and their applications, Applied Mechanics Reviews 55 (1) (2002) 1–34.
[388] 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.
[389] Jeremiah U Brackbill, Hans 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.
[390] Chenfanfu Jiang, Craig Schroeder, Andrew Selle, Joseph Teran, Alexey Stomakhin, The affine particle-in-cell method, ACM Transactions on Graphics (TOG) 34 (4) (2015) 51.
[391] Deborah Sulsky, Zhen Chen, Howard L Schreyer, A particle method for history-dependent materials, Computer methods in applied mechanics and engineering 118 (1-2) (1994) 179–196.
[392] Deborah Sulsky, Shi-Jian Zhou, Howard L Schreyer, Application of a particle-in-cell method to solid mechanics, Computer physics communications 87 (1-2) (1995) 236–252.
[393] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, Andrew Selle, A material point method for snow simulation, ACM Transactions on Graphics (TOG) 32 (4) (2013) 102.
[394] Johan Gaume, T Gast, J Teran, A van Herwijnen, C Jiang, Dynamic anticrack propagation in snow, Nature communications 9 (1) (2018) 3047.
[395] Daniel Ram, Theodore Gast, Chenfanfu Jiang, Craig Schroeder, Alexey Stomakhin, Joseph Teran, Pirouz Kavehpour, A material point method for viscoelastic fluids, foams and sponges, in: ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA), 2015.
[396] Yonghao Yue, Breannan Smith, Christopher Batty, Changxi Zheng, Eitan Grinspun, Continuum foam: A material point method for shear-dependent flows, ACM Transactions on Graphics (TOG) 34 (5) (2015) 160.
[397] Yu Fang, Minchen Li, Ming Gao, Chenfanfu Jiang, Silly rubber: an implicit material point method for simulating non-equilibrated viscoelastic and elastoplastic solids, ACM Transactions on Graphics (TOG) 38 (4) (2019) 118.
[398] Gergely Klar, Theodore Gast, Andre Pradhana, Chuyuan Fu, Craig Schroeder, Chenfanfu Jiang, Joseph Teran, Drucker-prager elastoplasticity for sand animation, ACM Transactions on Graphics (TOG) 35 (4) (2016) 103.
[399] Gilles Daviet, Florence Bertails-Descoubes, A semi-implicit material point method for the continuum simulation of granular materials, ACM Transactions on Graphics (TOG) 35 (4) (2016) 102.
[400] Yuanming Hu, Yu Fang, Ziheng Ge, Ziyin Qu, Yixin Zhu, Andre Pradhana, Chenfanfu 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.
[401] Stephanie Wang, Mengyuan Ding, Theodore F Gast, Leyi Zhu, Steven Gagniere, Chenfanfu Jiang, Joseph M Teran, Simulation and visualization of ductile fracture with the material point method, ACM Transactions on Graphics (TOG) 2 (2) (2019) 18.
[402] Joshuah Wolper, Yu Fang, Minchen Li, Jiecong Lu, Ming Gao, Chenfanfu Jiang, Cd-mpm: continuum damage material point methods for dynamic fracture animation, ACM Transactions on Graphics (TOG) 38 (4) (2019) 119.
[403] Chenfanfu Jiang, Theodore Gast, Joseph Teran, Anisotropic elastoplasticity for cloth, knit and hair frictional contact, ACM Transactions on Graphics (TOG) 36 (4) (2017) 152.
[404] Xuchen Han, Theodore F Gast, Qi Guo, Stephanie Wang, Chenfanfu Jiang, Joseph Teran, A hybrid material point method for frictional contact with diverse materials, ACM Transactions on Graphics (TOG) 2 (2) (2019) 17.
[405] Chuyuan Fu, Qi Guo, Theodore Gast, Chenfanfu Jiang, Joseph Teran, A polynomial particle-in-cell method, ACM Transactions on Graphics (TOG) 36 (6) (2017) 222.
[406] Alexey Stomakhin, Craig Schroeder, Chenfanfu Jiang, Lawrence Chai, Joseph Teran, Andrew Selle, Augmented mpm for phase-change and varied materials, ACM Transactions on Graphics (TOG) 33 (4) (2014) 138.
[407] Andre Pradhana Tampubolon, Theodore Gast, Gergely Klár, Chuyuan Fu, Joseph Teran, Chenfanfu Jiang, Ken Museth, Multi-species simulation of porous sand and water mixtures, ACM Transactions on Graphics (TOG) 36 (4) (2017) 105.
[408] Ming Gao, Andre Pradhana, Xuchen Han, Qi Guo, Grant Kot, Eftychios Sifakis, Chenfanfu Jiang, Animating fluid sediment mixture in particle-laden flows, ACM Transactions on Graphics (TOG) 37 (4) (2018) 149.
[409] John A Nairn, Material point method calculations with explicit cracks, Computer Modeling in Engineering and Sciences 4 (6) (2003) 649–664.
[410] 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.
[411] HL Schreyer, DL 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.
[412] Deborah Sulsky, Howard 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.
[413] Peng Huang, X Zhang, S Ma, HK 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.
[414] Wenqing Hu, Zhen 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.
[415] Allen R York, Deborah Sulsky, Howard L Schreyer, Fluid–membrane interaction based on the material point method, International Journal for Numerical Methods in Engineering 48 (6) (2000) 901–924.
[416] Samila Bandara, Kenichi Soga, Coupling of soil deformation and pore fluid flow using material point method, Computers and geotechnics 63 (2015) 199–214.
[417] James E Guilkey, James B Hoying, Jeffrey A Weiss, Computational modeling of multicellular constructs with the material point method, Journal of biomechanics 39 (11) (2006) 2074–2086.
[418] Peng HUANG, Material point method for metal and soil impact dynamics problems, Tsinghua University, 2010.
[419] Yu Fang, Yuanming Hu, Shi-Min Hu, Chenfanfu Jiang, A temporally adaptive material point method with regional time stepping, in: Computer Graphics Forum, 2018.
[420] SG Bardenhagen, EM Kober, The generalized interpolation material point method, Computer Modeling in Engineering and Sciences 5 (6) (2004) 477–496.
[421] Ming Gao, Andre Pradhana Tampubolon, Chenfanfu Jiang, Eftychios Sifakis, An adaptive generalized interpolation material point method for simulating elastoplastic materials, ACM Transactions on Graphics (TOG) 36 (6) (2017) 223.
[422] A Sadeghirad, Rebecca M Brannon, J Burghardt, A convected particle do-

main 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.
[423] Duan Z Zhang, Xia Ma, Paul T Giguere, Material point method enhanced by modified gradient of shape function, Journal of Computational Physics 230 (16) (2011) 6379–6398.
[424] Daniel S Bernstein, Robert Givan, Neil Immerman, Shlomo Zilberstein, The complexity of decentralized control of markov decision processes, Mathematics of operations research 27 (4) (2002) 819–840.
[425] Claudia V Goldman, Shlomo Zilberstein, Optimizing information exchange in cooperative multi-agent systems, in: International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2003.
[426] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller, Playing atari with deep reinforcement learning, arXiv preprint arXiv:1312.5602.
[427] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, Raul Vicente, Multiagent cooperation and competition with deep reinforcement learning, PloS one 12 (4) (2017) e0172395.
[428] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, Shimon Whiteson, Counterfactual multi-agent policy gradients, in: AAAI Conference on Artificial Intelligence (AAAI), 2018.
[429] Katie Atkinson, Trevor Bench-Capon, Peter McBurney, A dialogue game protocol for multi-agent argument over proposals for action, Autonomous Agents and Multi-Agent Systems 11 (2) (2005) 153–171.
[430] M David Sadek, Philippe Bretier, Franck Panaget, Artimis: Natural dialogue meets rational agency, in: International Joint Conference on Artificial Intelligence (IJCAI), 1997.
[431] Zilong Zheng, Wenguan Wang, Siyuan Qi, Song-Chun Zhu, Reasoning visual dialogs with structural and partial observations, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[432] Sainbayar Sukhbaatar, Rob Fergus, et al., Learning multiagent communication with backpropagation, in: Advances in Neural Information Processing Systems (NeurIPS), 2016.
[433] Igor Mordatch, Pieter Abbeel, Emergence of grounded compositional language in multi-agent populations, in: AAAI Conference on Artificial Intelligence (AAAI), 2018.
[434] Angeliki Lazaridou, Alexander Peysakhovich, Marco Baroni, Multi-agent cooperation and the emergence of (natural) language, in: International Conference on Learning Representations (ICLR), 2017.
[435] Serhii Havrylov, Ivan Titov, Emergence of language with multi-agent games: Learning to communicate with sequences of symbols, in: Advances in Neural Information Processing Systems (NeurIPS), 2017.
[436] Katrina Evtimova, Andrew Drozdov, Douwe Kiela, Kyunghyun Cho, Emergent language in a multi-modal, multi-step referential game, arXiv preprint arXiv:1705.10369.
[437] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, Stephen Clark, Emergence of linguistic communication from referential games with symbolic and pixel input, in: International Conference on Learning Representations (ICLR), 2018.
[438] Kyle Wagner, James A Reggia, Juan Uriagereka, Gerald S Wilkinson, Progress in the simulation of emergent communication and language, Adaptive Behavior 11 (1) (2003) 37–69.
[439] Rasmus Ibsen-Jensen, Josef Tkadlec, Krishnendu Chatterjee, Martin A Nowak, Language acquisition with communication between learners, Journal of The Royal Society Interface 15 (140) (2018) 20180073.
[440] Laura Graesser, Kyunghyun Cho, Douwe Kiela, Emergent linguistic phenomena in multi-agent communication games, arXiv preprint arXiv:1901.08706.
[441] Emmanuel Dupoux, Pierre Jacob, Universal moral grammar: a critical appraisal, Trends in Cognitive Sciences 11 (9) (2007) 373–378.
[442] John Mikhail, Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment, Cambridge University Press, 2011.
[443] Max Kleiman-Weiner, Rebecca Saxe, Joshua B Tenenbaum, Learning a commonsense moral theory, cognition 167 (2017) 107–123.
[444] PR Blake, K McAuliffe, J Corbit, TC Callaghan, O Barry, A Bowie, L Kleutsch, KL Kramer, E Ross, H Vongsachang, et al., The ontogeny of fairness in seven societies, Nature 528 (7581) (2015) 258.
[445] Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, In search of homo economicus: behavioral experiments in 15 small-scale societies, American Economic Review 91 (2) (2001) 73–78.
[446] Bailey R House, Joan B Silk, Joseph Henrich, H Clark Barrett, Brooke A Scelza, Adam H Boyette, Barry S Hewlett, Richard McElreath, Stephen Laurence, Ontogeny of prosocial behavior across diverse societies, Proceedings of the National Academy of Sciences (PNAS) 110 (36) (2013) 14586–14591.
[447] Jesse Graham, Peter Meindl, Erica Beall, Kate M Johnson, Li Zhang, Cultural differences in moral judgment and behavior, across and within societies, Current Opinion in Psychology 8 (2016) 125–130.
[448] Thomas Hurka, Virtue, vice, and value, Oxford University Press, 2000.
[449] John Rawls, A theory of justice, Harvard university press, 1971.
[450] Jonathan Haidt, The new synthesis in moral psychology, science 316 (5827) (2007) 998–1002.
[451] J Kiley 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.
[452] Richard Kim, Max Kleiman-Weiner, Andrés Abeliuk, Edmond Awad, Sohan Dsouza, Joshua B Tenenbaum, Iyad Rahwan, A computational model of commonsense moral decision making, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2018.
[453] Max Kleiman-Weiner, Tobias Gerstenberg, Sydney Levine, Joshua B Tenenbaum, Inference of intention and permissibility in moral decision making., in: Annual Meeting of the Cognitive Science Society (CogSci), 2015.
[454] Keith J Holyoak, Paul Thagard, The analogical mind, American psychologist 52 (1) (1997) 35.
[455] Mary B Hesse, Models and analogies in science, Notre Dame University Press, 1966.
[456] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, Jeff Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems (NeurIPS), 2013.
[457] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.
[458] Patricia A Carpenter, Marcel A Just, Peter Shell, What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test, Psychological review 97 (3) (1990) 404.
[459] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, Devi Parikh, Vqa: Visual question answering, in: International Conference on Computer Vision (ICCV), 2015.
[460] R E Snow, Patrick Kyllonen, B Marshalek, The topography of ability and learning correlations, Advances in the psychology of human intelligence (1984) 47–103.
[461] Susanne M Jaeggi, Martin Buschkuehl, John Jonides, Walter J Perrig, Improving fluid intelligence with training on working memory, Proceedings of the National Academy of Sciences (PNAS) 105 (19) (2008) 6829–6833.
[462] Gordon H Bower, A contrast effect in differential conditioning, Journal of Experimental Psychology 62 (2) (1961) 196.
[463] Donald R Meyer, The effects of differential rewards on discrimination reversal learning by monkeys, Journal of Experimental Psychology 41 (4) (1951) 268.
[464] Allan M Schrier, Harry F Harlow, Effect of amount of incentive on discrimination learning by monkeys, Journal of comparative and physiological psychology 49 (2) (1956) 117.
[465] Robert M Shapley, Jonathan D Victor, The effect of contrast on the transfer properties of cat retinal ganglion cells, The Journal of physiology 285 (1) (1978) 275–298.
[466] Reed 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.
[467] Abram Amsel, Frustrative nonreward in partial reinforcement and discrimination learning: Some recent history and a theoretical extension, Psychological review 69 (4) (1962) 306.
[468] James J Gibson, Eleanor J Gibson, Perceptual learning: Differentiation or enrichment?, Psychological review 62 (1) (1955) 32.
[469] James J Gibson, The ecological approach to visual perception: classic edition, Psychology Press, 2014.
[470] Richard Catrambone, Keith J Holyoak, Overcoming contextual limitations on problem-solving transfer, Journal of Experimental Psychology: Learning, Memory, and Cognition 15 (6) (1989) 1147.
[471] Dedre Gentner, Virginia Gunn, Structural alignment facilitates the noticing of differences, Memory & Cognition 29 (4) (2001) 565–577.
[472] Rubi Hammer, Gil Diesendruck, Daphna Weinshall, Shaul Hochstein, The development of category learning strategies: What makes the difference?, Cognition 112 (1) (2009) 105–119.
[473] Mary L Gick, Katherine Paterson, Do contrasting examples facilitate schema acquisition and analogical transfer?, Canadian Journal of Psychology/Revue canadienne de psychologie 46 (4) (1992) 539.
[474] Etsuko Haryu, Mutsumi Imai, Hiroyuki Okada, Object similarity bootstraps young children to action-based verb extension, Child Development 82 (2) (2011) 674–686.
[475] Linsey Smith, Dedre Gentner, The role of difference-detection in learning contrastive categories, in: Annual Meeting of the Cognitive Science Society (CogSci), 2014.
[476] Dedre Gentner, Structure-mapping: A theoretical framework for analogy, Cognitive science 7 (2) (1983) 155–170.
[477] Dedre Gentner, Arthur B Markman, Structural alignment in comparison: No difference without similarity, Psychological science 5 (3) (1994) 152–158.
[478] Daniel L Schwartz, Catherine C Chase, Marily A Oppezzo, Doris 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.
[479] Chi Zhang, Baoxiong Jia, Feng Gao, Yixin Zhu, Hongjing Lu, Song-Chun Zhu, Learning perceptual inference by contrasting, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.