Dark, Beyond Deep: A Paradigm Shift to Cognitive AI with Humanlike Common Sense

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 in deep learning is essentially based on a "big data for small tasks" paradigm, under which massive amounts of data are used to train a classifier for a single narrow task. In this paper, we call for a shift that flips this paradigm upside down. Specifically, we propose a "small data for big tasks" paradigm, wherein a single artificial intelligence (AI) system is challenged to develop "common sense," enabling it to solve a wide range of tasks with little training data. We illustrate the potential power of this new paradigm by reviewing models of common sense that synthesize recent breakthroughs in both machine and human vision. We identify functionality, physics, intent, causality, and utility (FPICU) as the five core domains of cognitive AI with humanlike common sense. When taken as a unified concept, FPICU is concerned with the questions of "why" and "how," beyond the dominant "what" and "where" framework for understanding vision. They are invisible in terms of pixels but nevertheless drive the creation, maintenance, and development of visual scenes. We therefore coin them the "dark matter" of vision. Just as our universe cannot be understood by merely studying observable matter, we argue that vision cannot be understood without studying FPICU. We demonstrate the power of this perspective to develop cognitive AI systems with humanlike common sense by showing how to observe and apply FPICU with little training data to solve a wide range of challenging tasks, including tool use, planning, utility inference, and social learning. In summary, we argue that the next generation of AI must embrace "dark" humanlike common sense for solving novel tasks.

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

1. A Call for a Paradigm Shift in Vision and AI

Computer vision is 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 at "what" is "where." Here, "what" refers to object recognition (object vision), and "where" denotes three-dimensional (3D) reconstruction and object localization (spatial vision) [2]. Such a definition corresponds to two pathways in the human brain: (i) the ventral pathway for categorical recognition of objects and scenes, and (ii) the dorsal pathway for the reconstruction of depth and shapes, scene layout, visually guided actions, and so forth. This paradigm guided the geometry-based approaches to computer vision of the 1980s-1990s, and the appearance-based methods of the past 20 years.

Over the past several years, progress has been made in object detection and localization with the rapid advancement of deep neural networks (DNNs), fueled by hardware accelerations and the availability of massive sets of labeled data. However, we are still far from solving computer vision or real machine intelligence; the inference and reasoning abilities of current computer vision systems are narrow and highly specialized, require large sets of labeled training data designed for special tasks, and lack a general *understanding* of common facts—that is, facts that are obvious to the average human adult—that describe how our physical and social worlds work. To fill in the gap between modern computer vision and human vision, we must find a broader perspective from which to model and reason about the missing dimension, which is humanlike common sense.

This state of our understanding of vision is analogous to what has been observed in the fields of cosmology and astrophysicists. In the 1980s, physicists proposed what is now the standard cosmology model, in which the mass-energy observed by the electromagnetic spectrum accounts for less than 5% of the universe; the rest of the universe is dark matter (23%) and dark energy (72%)[1]. The properties and characteristics of dark matter and dark energy cannot be observed and must be reasoned from the visible mass-energy using a sophisticated model. Despite their invisibility, however, dark matter and energy 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 community and beyond, of the missing dimensions and the potential benefits of joint representation and joint inference. We argue that humans can make rich inferences

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

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

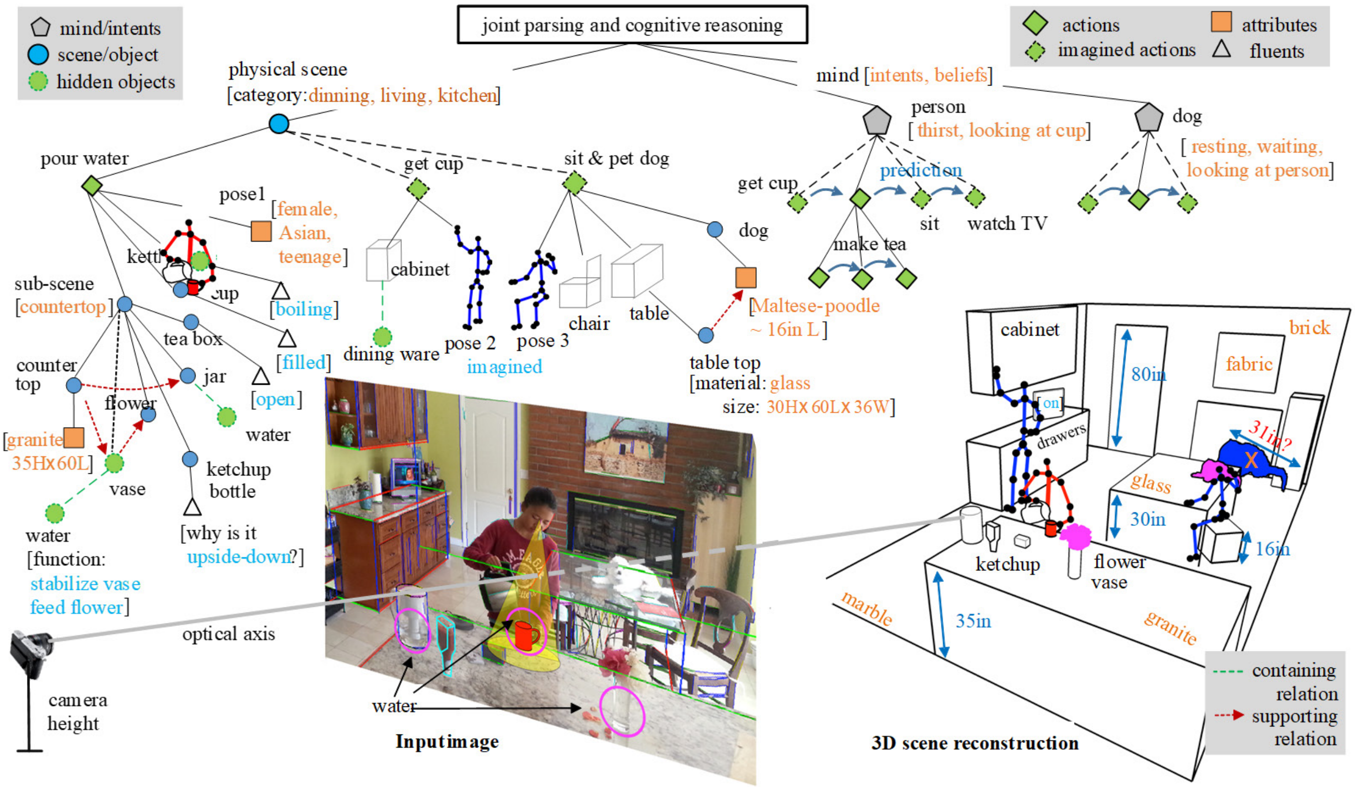


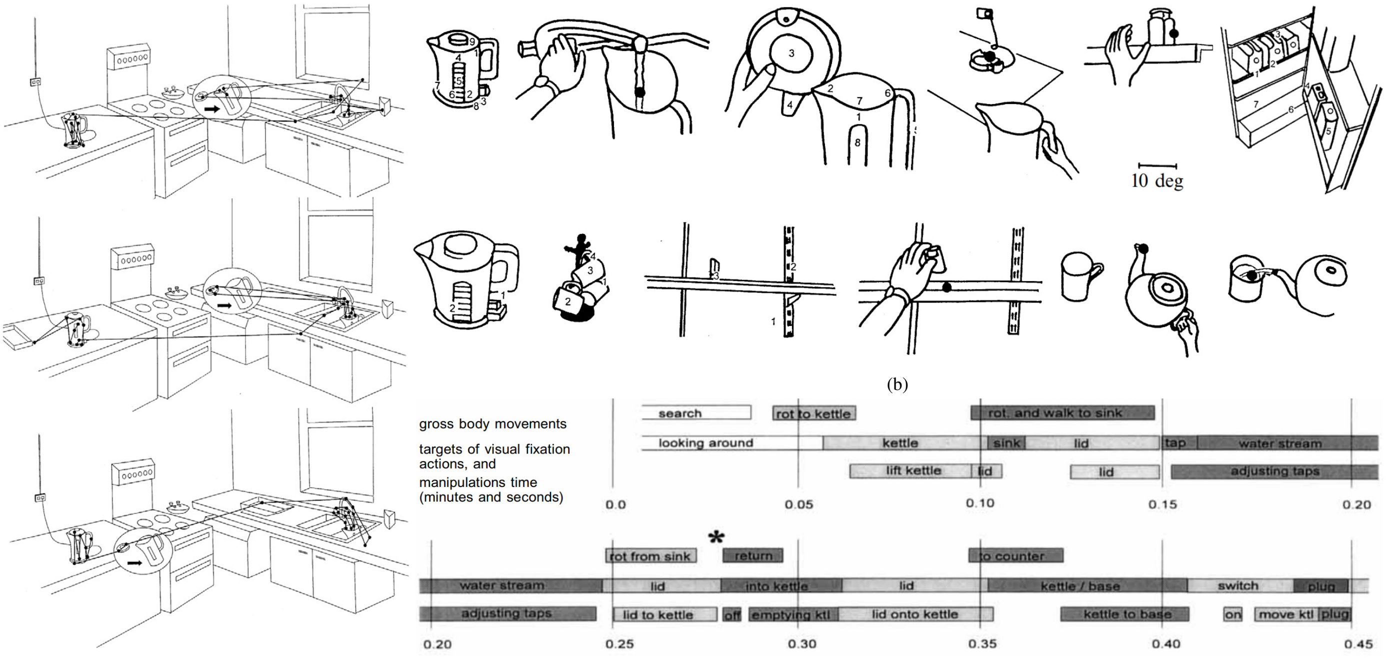
Figure 1: An example of in-depth understanding of a scene or event through joint parsing and cognitive reasoning. From a single image, a computer vision system should be able to jointly (i) reconstruct the 3D scene; (ii) estimate camera parameters, materials, and illumination; (iii) parse the scene hierarchically with attributes, fluents, and relationships; (iv) reason about the intentions and beliefs of agents (*e.g.*, the human and dog in this example); (v) predict their actions in time; and (vi) recover invisible elements such as water, latent object states, and so forth. We, as humans, can effortlessly (i) predict that water is about to come out of the kettle; (ii) reason that the intent behind putting the ketchup bottle upside down is to utilize gravity for easy use; and (iii) see that there is a glass table, which is difficult to detect with existing computer vision methods, under the dog; without seeing the glass table, parsing results would violate the laws of physics, as the dog would appear to be floating in midair. These perceptions can only be achieved by reasoning about unobservable factors in the scene not represented by pixels, requiring us to build an AI system with humanlike core knowledge and common sense, which are largely missing from current computer vision research. H: height; L: length; W: width. 1 in = 2.54 cm.

from sparse and high-dimensional data, and achieve deep understanding from a single picture, because we have common yet visually imperceptible knowledge that can never be understood just by asking "what" and "where." Specifically, human-made objects and scenes are designed with latent functionality, determined by the unobservable laws of physics and their downstream causal relationships; consider how our understanding of water's flow from of a kettle, or our knowledge that a transparent substance such as glass can serve as a solid table surface, tells us what is happening in Fig. 1. Meanwhile, human activities, especially social activities, are governed by causality, physics, functionality, social intent, and individual preferences and utility. In images and videos, many entities (*e.g.*, functional objects, fluids, object fluents, and intent) and relationships (*e.g.*, causal effects and physical supports) are impossible to detect by most of the existing approaches considering appearance alone; these latent factors are not represented in pixels. Yet they are pervasive and govern the placement and motion of the visible entities that are relatively easy for current methods to detect.

These invisible factors are largely missing from recent computer vision literature, in which most tasks have been converted into classification problems, empowered by large-scale annotated data and end-to-end training using neural networks. This is what we call the "big data for small tasks" paradigm of computer vision and AI.

In this paper, we aim to draw attention to a promising new direction, where consideration of "dark" entities and relationships is incorporated into vision and AI research. By reasoning about the unobservable factors beyond visible pixels, we could approximate humanlike common sense, using limited data to achieve generalizations across a variety of tasks. Such tasks would include a mixture of both classic "what and where" problems (*i.e.*, classification, localization, and reconstruction), and "why, how, and what if" problems, including but not limited to causal reasoning, intuitive physics, learning functionality and affordance, intent prediction, and utility learning. 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 meanings of "latent" or "hidden" used in vision and probabilistic modeling; "darkness" is a measure of the relative difficulty of classifying an entity or inferring about a relationship based on how much invisible common



(a) (b) (c)

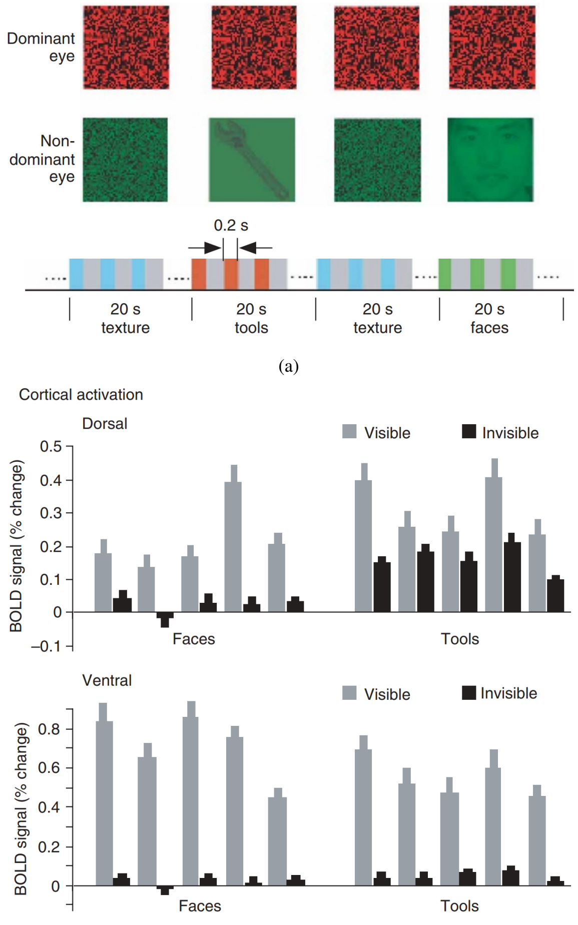
Figure 2: Even for as "simple" a task as making a cup of tea, a person can make use of his or her single vision system to perform a variety of subtasks in order to achieve the ultimate goal. (a) Record of the visual fixations of three different subjects performing the same task of making a cup of tea in a small rectangular kitchen; (b) examples of fixation patterns drawn from an eye-movement videotape; (c) a sequence of visual and motor events during a tea-making session. Rot: rotate; ktl: kettle. Reproduced from Ref. [3] with permission of SAGE Publication, © 1999.

sense needed beyond the visible appearance or geometry. Entities can fall on a continuous spectrum of "darkness"—from objects such as a generic human face, which is relatively easy to recognize based on its appearance, and is thus considered "visible," to functional objects such as chairs, which are challenging to recognize due to their large intraclass variation, and all the way to entities or relationships that are impossible to recognize through pixels. In contrast, the functionality of the kettle is "dark;" through common sense, a human can easily infer that there is liquid inside it. The position of the ketchup bottle could also be considered "dark," as the understanding of typical human intent lets us understand that it has been placed upside down to harness gravity for easy dispensing.

The remainder of this paper starts by revisiting a classic view of computer vision in terms of "what" and "where" in Section 2, in which we show that the human vision system is essentially task-driven, with its representation and computational mechanisms rooted in various tasks. In order to use "small data" to solve "big tasks," we then identify and review five crucial axes of visual common sense: **F**unctionality, **P**hysics, perceived **I**ntent, **C**ausality, and **U**tility (FPICU). Causality (Section 3) is the basis for intelligent understanding. The application of causality (*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 humans use when they interact with it, performing appropriate actions to change the world in service of activities. When considering social interactions beyond the physical world, humans need to further infer intent (Section 6) in order to understand other humans' 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 relationships" indeed support various visual tasks beyond just classification. We summarize and discuss our perspectives in Section 8, arguing that it is crucial for the future of AI to master these essential unseen ingredients, rather than only increasing the performance and complexity of data-driven approaches.

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

What should a vision system afford the agent it serves? From a biological perspective, the majority of living creatures use a *single* (with multiple components) vision system to perform *thousands* of tasks. This contrasts with the dominant contemporary stream of thought in computer vision research, where a single model is designed specifically for a single task. In the literature, this organic paradigm of generalization, adaptation, and transfer among various tasks is referred to as task-centered vision [4]. In the kitchen shown in Fig. 2 [3], even a task as simple as making a cup of coffee consists of multiple subtasks, including finding objects (object recognition), grasping objects (object manipulation), finding milk in the refrigerator, and adding sugar (task planning). Prior research has shown that a person can finish making a cup of coffee within 1 min by utilizing a single vision system to facilitate the performance of a variety of subtasks [3].



(a)

(b)

Figure 3: Cortical responses to invisible objects in the human dorsal and ventral pathways. (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 responsive to tools, but not to faces, while neither tools or faces evoked much activation in the ventral area. BOLD: blood oxygen level-dependent. Reproduced from Ref. [5] with permission of Nature Publishing Group, © 2005.

Neuroscience studies suggest similar results, indicating that the human vision system is far more capable than any existing computer vision system, and goes beyond merely memorizing patterns of pixels. For example, Fang and He [5] showed that recognizing a face inside an image utilizes a different mechanism from recognizing an object that can be manipulated as a tool, as shown in Fig. 3; indeed, their results show that humans may be even more visually responsive to the appearance of tools than to faces, driving home how much reasoning about how an object can help perform tasks is ingrained in visual intelligence. Other studies [6] also support the similar conclusion that images of tools "potentiate" actions, even when overt actions are not required. Taken together, these results indicate that our biological vision system possesses a mechanism for perceiving object functionality (*i.e.*, how an object can be manipulated as a tool) that is independent of the mechanism governing face recognition (and recognition of other objects). All these findings call for a quest to discover the mechanisms of the human vision system and natural intelligence.

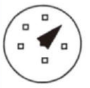
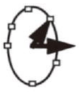
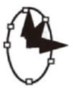
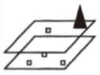
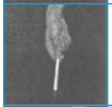
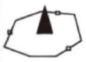
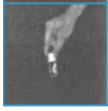
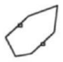
Grasp strategy	Required functional capabilities	Representation
	 ~Center ~Radius	Superquadrics
	 ~Center ~Radius ~Axis direction	Generalized cylinder
	 ~Center ~Radius ~Axis direction ~Pulling direction	Superquadrics + pulling direction
	 Orientation Position of two planes Width	Two parallel planes (geometric 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. Reproduced from Ref. [7] with permission of IEEE, © 1992.

2.1. "What": Task-centered Visual Recognition

The human brain can grasp the "gist" of a scene in an image within 200 ms, as observed by Potter in the 1970s [8, 9], and by Schyns and Oliva [10] and Thorpe *et al.* [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 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 good performances on scene categorization in terms of recognition accuracy in publicly available datasets, a recent large-scale neuroscience study [23] has shown that current DNNs cannot account for the image-level behavior patterns of primates (both humans and monkeys), calling attention to the need for more precise accounting for the neural mechanisms underlying primate object vision. Furthermore, data-driven approaches have led the focus of scene categorization research away from an important determinant of visual information—the categorization task itself [24, 25]. Simultaneously, these approaches have left 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 "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 one's own kitchen (Fig. 5) [22]. As shown in Ref. [22], scene categorization and the information-gathering process are constrained by these categorization tasks [26, 27], suggesting a bidirectional interplay between the visual input

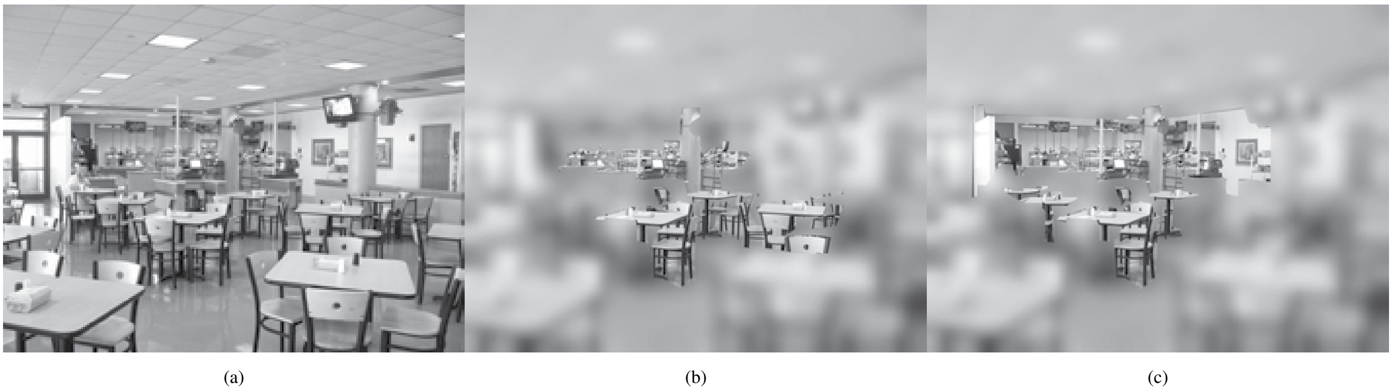


Figure 5: The experiment presented in Ref. [22], demonstrating the diagnostically driven, bidirectional interplay between top-down and bottom-up information for the categorization of scenes at specific hierarchical levels. (a) Given the same input image of a scene, subjects will show different gaze patterns if they are asked to categorize the scene at (b) a basic level (*e.g.*, restaurant) or (c) a subordinate level (*e.g.*, cafeteria), indicating a task-driven nature of scene categorization. Reproduced from Ref. [22] with permission of the authors, © 2014.

and the viewer's needs/tasks [25]. Beyond scene categorization, similar phenomena were also observed in facial recognition [28].

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

2.2. *"Where": Constructing 3D Scenes as a Series of Tasks*

In the literature, approaches to 3D machine vision have assumed that the goal is to build an accurate 3D model of the scene from the camera/observer's perspective. These structure-from-motion (SfM) and simultaneous localization and mapping (SLAM) methods [29] have been the prevailing paradigms in 3D scene reconstruction. In particular, scene reconstruction from a single two-dimensional (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 enable agents to perform tasks by generating the best possible configuration in terms of functionality, physics, and object relationships. This line of work has mostly been studied separately from recognition and semantics until recently [31, 32, 33, 34, 35, 36, 37, 38]; see Fig. 6 [36] for an example.

The idea of reconstruction as a "cognitive map" has a long history [39]. 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 [40, 41]. In fact, multiple experimental studies do not countenance global metric representations [42, 43, 44, 45, 46, 47]; human vision is error-prone and distorted in terms of localization [48, 49, 50, 51, 52]. In a case study, Glennerster *et al.* [53] demonstrated an astonishing lack of sensitivity on the part of observers to dramatic changes in the scale of the environment around a moving observer performing various tasks.

Among all the recent evidence, grid cells are perhaps the most well-known discovery to indicate the non-necessity of precise 3D reconstruction for vision tasks [54, 55, 56]. Grid cells encode a cognitive representation of Euclidean space, implying a different mechanism for 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 [57], but is also found in mice [58, 59], bats [60], and other animals. Gao *et al.* [61] and Xie *et al.* [62] proposed a representational model for grid cells, in which the 2D self-position of an 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 showed that view-based methods actually perform better than 3D reconstruction-based methods in certain human navigation tasks [63].

Despite these discoveries, how we navigate complex environments while remaining able at all times to return to an original location (*i.e.*, homing) remains a mystery in biology and neuroscience. Perhaps a recent study from Vuong *et al.* [64] providing evidence for the task-dependent representation of space can shed some light. Specifically, in this experiment, participants made large, consistent pointing errors that were poorly explained by any single 3D representation. Their study suggests that the mechanism for maintaining visual directions for reaching unseen targets is neither based on a stable 3D model of a scene nor a distorted one; instead, participants seemed to form a flat and task-dependent representation.

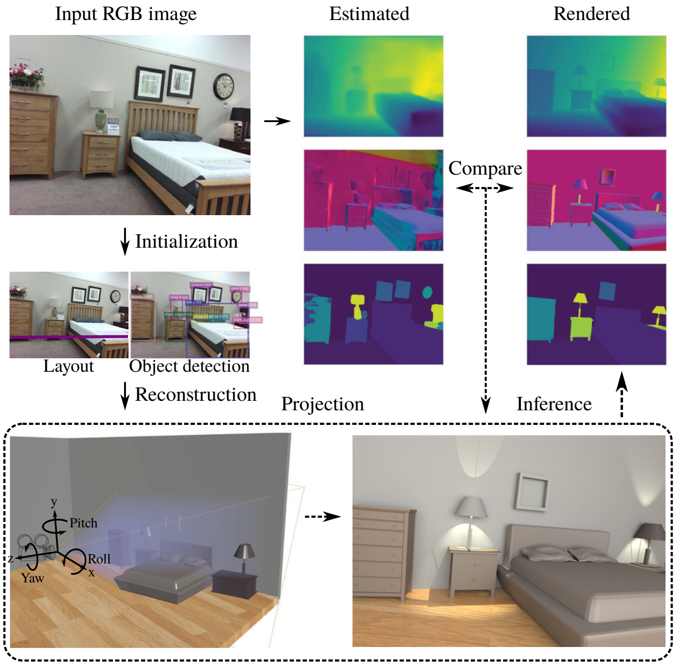


Figure 6: Illustration of 3D indoor scene parsing and reconstruction in an analysis-by-synthesis fashion [36]. A 3D representation is initialized by individual vision tasks (*e.g.*, object detection, 2D layout estimation). A joint inference algorithm compares the differences between the rendered normal, depth, and segmentation maps and the ones estimated directly from the input RGB image, and adjusts the 3D structure iteratively. Reproduced from Ref. [36] with permission of Springer, © 2018.

2.3. Beyond "What" and "Where": Towards Scene Understanding with Humanlike Common Sense

Psychological studies have shown that human visual experience is much richer than "what" and "where." As early as infancy, humans quickly and efficiently perceive causal relationships (*e.g.*, perceiving that object A launches object B) [65, 66], agents and intentions (*e.g.*, understanding that one entity is chasing another) [67, 68, 69], and the consequences of physical forces (*e.g.*, predicting that a precarious stack of rocks is about to fall in a particular direction) [70, 71]. Such physical and social concepts can be perceived from both media as rich as videos [72] and much sparser visual inputs [73, 74]; see examples in Fig. 11.

To enable an artificial agent with similar capabilities, we call for joint reasoning algorithms on a joint representation that integrates (i) the "visible" traditional recognition and categorization of objects, scenes, actions, events, and so forth; and (ii) the "dark" higher level concepts of fluent, causality, physics, functionality, affordance, intentions/goals, utility, and so forth. These concepts can in turn be divided into five axes: fluent and perceived causality, intuitive physics, functionality, intentions and goals, and utility and preference, described below.

2.3.1. Fluent and Perceived Causality

A *fluent*, which is a concept coined and discussed by Isaac Newton [75] and Maclaurin [76], respectively, and adopted by AI and commonsense reasoning [77, 78], refers to a transient state of an object that is time-variant, such as a cup being empty or filled, a door being locked, a car blinking to signal a left turn, and a telephone ringing; see Fig. 7 for other examples of "dark" fluents in images. Fluents are linked to perceived causality [79] in the psychology literature. Even infants with limited 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 events [80].

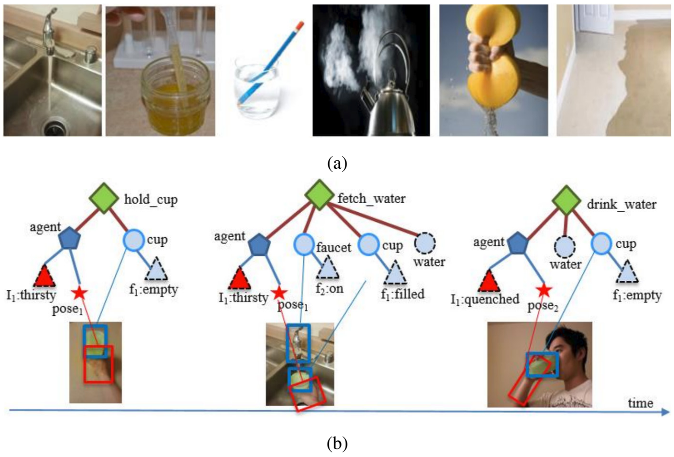


Figure 7: Water and other clear fluids play important roles in a human's daily life, but are barely detectable in images. (a) Water causes only minor changes in appearance; (b) the "dark" entities of water, fluents (here, a cup and faucet, represented by triangles), and the intention of a human are shown in dashed nodes. The actions (diamonds) involve agents (pentagons) and cups (objects in circles).

Fluents and perceived causality are different from the visual *attributes* [82, 83] of objects. The latter are permanent over the course of observation; for example, the gender of a person in a short video clip should be an attribute, not a fluent. Some fluents are visible, but many are "dark." Human cognition has the innate capability (observed in infants) [80] and strong inclination to perceive the *causal effects* between *actions* and *changes of fluents*; for example, realizing that flipping a switch causes a light to turn on. To recognize the change in an object caused by an action, one must be able to perceive and evaluate the state of the object's changeable characteristics; thus, perceiving fluents, such as whether the light switch is set to the up or down position, is essential for recognizing actions and understanding events as they unfold. Most vision research on action recognition has paid a great deal of attention to the position, pose, and movement of the human body in the process of activities such as walking, jumping, and clapping, and to human-object interactions such as drinking and smoking [84, 85, 86, 87]; but most daily actions, such as opening a door, are defined by cause and effect (a door's fluent changes from "closed" to "open," regardless of how it is opened), rather than by the human's position, movement, or spatial-temporal features [88, 89]. Similarly, actions such as putting on clothes or setting up a tent cannot be defined simply by their appearance features; their complexity demands causal reasoning to be understood. Overall, the status of a scene can be viewed as a collection of fluents that *record the history of actions*. Nevertheless, fluents and causal reasoning have not yet been systematically studied in machine vision, despite their ubiquitous presence in images and videos.

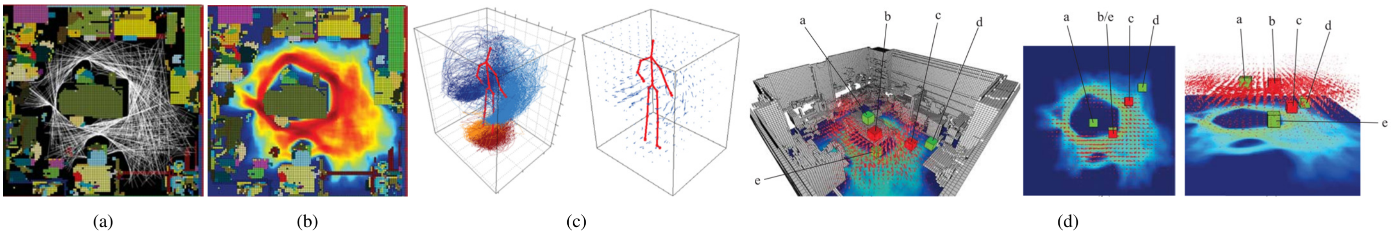


Figure 8: Inferring the potential for objects to fall from human actions and natural disturbances. (a) The imagined human trajectories; (b) the distribution of primary motion space; (c) the secondary motion field; (d) the integrated human action field, built by integrating primary motions with secondary motions. The five objects **a-e** are typical cases in the disturbance field: The objects **b** on the edge of a table and **c** along the pathway exhibit greater disturbance (in the form of accidental collisions) than other objects such as **a** in the center of the table, **e** below the table, and **d** in a concave corner of the room. Reproduced from Ref. [81] with permission of IEEE, © 2014.

2.3.2. Intuitive Physics

Psychology studies suggest that approximate Newtonian principles underlie human judgments about dynamics and stability [90, 91]. Hamrick *et al.* [71] and Battaglia *et al.* [70] showed that the knowledge of Newtonian principles and probabilistic representations is generally applied in human physical reasoning, and that an intuitive physical model is an important aspect of human-level complex scene understanding. Other studies have shown that humans are highly sensitive to whether objects in a scene violate certain understood physical relationships or appear to be physically unstable [92, 93, 94, 95, 96].

Invisible physical fields govern the layout and placement of objects in a human-made scene. By human design, objects should be physically stable and safe with respect to gravity and various other potential disturbances [97, 81, 98], such as an earthquake, a gust of wind, or the actions of other humans. Therefore, any 3D scene interpretation or parsing (*e.g.*, object localization and segmentation) must be physically plausible [97, 81, 98, 99, 36, 100]; see Fig. 8. This observation sets useful constraints to scene understanding and is important for robotics applications [81]. For example, in a search-and-rescue mission at a disaster-relief site, a robot must be able to reason about the stability of various objects, as well as about which objects are physically supporting which other objects, and then use this information to move cautiously and avoid creating dangerous new disturbances.

2.3.3. Functionality

Most human-made scenes are designed to serve multiple human functions, such as sitting, eating, socializing, and sleeping, and to satisfy human needs with respect to those functions, such as illumination, temperature control, and ventilation. These functions and needs are invisible in images, but shape the scene's layout [101, 34], its geometric dimensions, the shape of its objects, and the selection of its materials.

Through functional magnetic resonance imaging (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 [102]. Concepts in the human mind are not only represented by prototypes—that is, exemplars as in current computer vision and machine learning approaches—but also by functionality [80].

2.3.4. Intentions and Goals

Cognitive studies [103] show that humans have a strong inclination to interpret events as a series of goals driven by the intentions of agents. Such a teleological stance inspired various models in the cognitive literature for intent estimation as an inverse planning problem [104, 105].

We argue that intent can be treated as the transient status of agents (humans and animals), such as being "thirsty," "hungry," or "tired." They are similar to, but more complex than, the fluents of objects, and come with the following characteristics: (i) They are hierarchically organized in a sequence of goals and are the main factors driving actions and events in a scene. (ii) They are completely "dark," that is, not represented by pixels. (iii) Unlike the instant change of fluents in response to actions, intentions are often formed across long spatiotemporal ranges. For example, in Fig. 9 [72], when a person is hungry and sees a food truck in the courtyard, the person decides (intends) to walk to the truck.

During this process, an attraction relationship 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 or repulsion fields (*e.g.*, foul odor, or grass to avoid stepping on), which are completely "dark." The trajectory of a person with a certain intention moving through these fields 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 intentions and goals will be crucial for the following vision and cognition tasks: (i) early event and trajectory prediction [106]; (ii) discovery of the invisible attractive/repulsive fields of objects and recognizing their functions by analyzing human trajectories [72]; (iii) understanding of scenes by function and activity [26], where the attraction fields are longer range in a scene than the functionality maps [27, 107] and affordance maps [108, 109, 110] studied in recent literature; (iv) understanding multifaceted relationships among a group of people and their functional roles [111, 112, 113]; and (v) understanding and inferring the mental states of agents [114, 115].

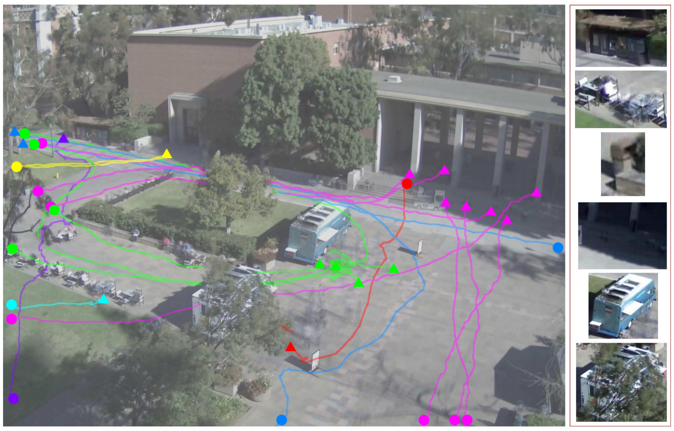


Figure 9: People's trajectories are color-coded to indicate their shared destination. The triangles denote destinations, and the dots denote start positions; *e.g.*, people may be heading toward the food truck to buy food (green), or to the vending machine to quench thirst (blue). Due to low resolution, poor lighting, and occlusions, objects at the destinations are very difficult to detect based only on their appearance and shape. Reproduced from Ref. [72] with permission of IEEE, © 2018.

2.3.5. 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 agents make near-optimal choices to minimize the cost of certain tasks; that is, we can assume there is no deception or pretense. This is known as the rational choice theory; that is, 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 that each agent *truthfully* reports its preferences; see Ref. [116] for a short introductory survey. Building computational models for human utility can be traced back to the English philosopher Jeremy Bentham, and to his works on ethics known as utilitarianism [117].

By observing a rational person's behavior and choices, it is possible to reverse-engineer their reasoning and learning process, and estimate their values. Utility, or values, are also used in the field of AI in planning schemes such as the Markov decision process (MDP), and are often associated with the states of a task. However, in the literature of the MDP, "value" is not a reflection of true human preference and, inconveniently, is tightly dependent on the agent's actions [118]. We argue that such utility-driven learning could be more invariant than traditional supervised training for computer vision and AI.

2.3.6. Summary

Despite their apparent differences at first glance, the five FPICU domains interconnect in ways that are theoretically important. These interconnections include the following characteristics: (i) The five FPICU domains usually do not easily project onto explicit visual features; (ii) most of the existing computer vision and AI algorithms are neither competent in these domains nor (in most cases) applicable at all; and (iii) human vision is nevertheless highly efficient in these domains, and human-level reasoning often builds upon prior knowledge and capability with FPICU.

We argue that the incorporation of these five key elements would advance a vision or AI system in at least three aspects:

1. Generalization. As a higher level representation, the FPICU concept tends to be globally invariant across the entire human living space. Therefore, knowledge learned in one scene can be transferred to novel situations.
2. Small sample learning. 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 learning of FPICU, which is more consistent and noise-free across different domains and data sources, is possible even without "big data."
3. Bidirectional inference. Inference with FPICU requires the combination of top-down inference based on abstract knowledge and bottom-up inference based on visual pattern. This means that systems would both continue to make data-driven inferences from the observation of visible, pixel-represented scene aspects, as they do today, and make inferences based on FPICU understanding. These two processes can feed on each other, boosting overall system performance.

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 [120, 121, 122]. People have innate assumptions about causes, and causal reasoning can be activated almost automatically and irresistibly [123, 124]. In our opinion, causality is the foundation of the other four FPICU elements (functionality, physics, intent, and utility). For example, an agent must be able to reason about the causes of others' behavior in order to understand their intent and understand the likely effects of their own actions to use functional objects appropriately. To a certain degree, much of human understanding depends on the ability to comprehend causality. Without understanding what causes an action, it is very difficult to consider what may happen next and respond effectively.

In this section, we start with a brief review of the causal perception and reasoning literature in psychology, followed by a review of a parallel stream of work in statistical learning. We conclude the section with case studies of causal learning in computer vision and 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, Chen and Scholl [125] showed that the human visual system can *perceive* causal history through commonsense visual reasoning, and can represent objects in terms of their inferred underlying causal history—essentially representing shapes by wondering about "how they got to be that

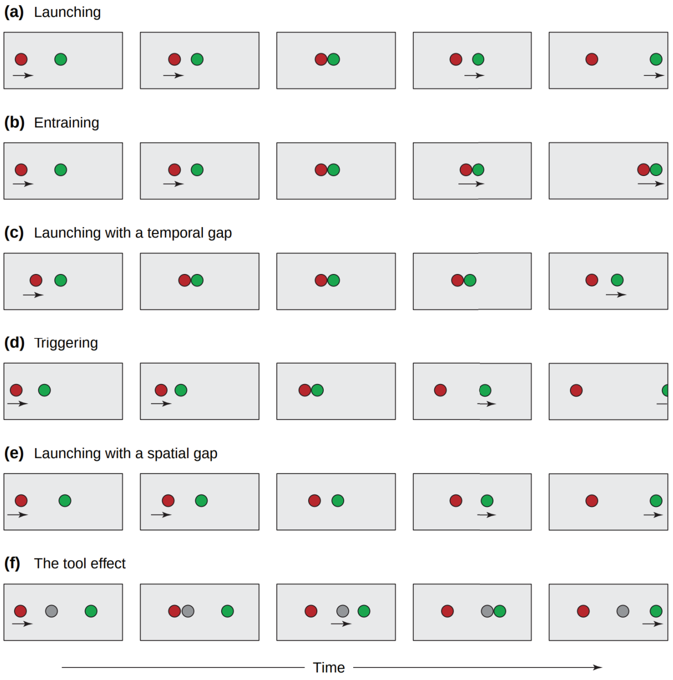
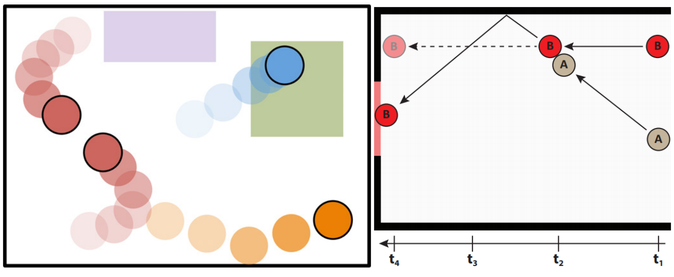

(a) Launching
(b) Entraining
(c) Launching with a temporal gap
(d) Triggering
(e) Launching with a spatial gap
(f) The tool effect
Time

Figure 10: Examples of some of Michotte's basic demonstrations of perceptual causality, regarding the perception of two objects, A and B (here shown as red and green circles, respectively). (a) The launching effect; (b) the entraining effect, wherein A seems to carry B along with it; (c) the launching effect is eliminated 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 eliminated by adding a spatial gap between A's final position and B's initial position; (f) the tool effect, wherein an intermediate item (gray circle) seems merely a tool by which A causes the entire motion sequence. These are some of the many cause-effect relationships between objects that humans understand intuitively, and that AI must learn to recognize. Reproduced from Ref. [119] with permission of Elsevier Science Ltd., © 2000.

way." Inherently, causal events cannot be directly interpreted merely from vision; they must be interpreted by an agent that understands the distal world [126].

Early psychological work focused on an associative mechanism as the basis for human causal learning and reasoning [127]. During this time, the Rescorla-Wagner model was used to explain how humans (and animals) build expectations using the cooccurrence of perceptual stimuli [128]. However, more recent studies have shown that human causal learning is a rational Bayesian process [126, 129, 130] involving the acquisition of *abstract* causal structure [131, 132] and strength values for cause-effect relationships [133].

The perception of causality was first systematically studied by the psychologist Michotte [79] through observation of one billiard ball (A) hitting another (B); see Fig. 10 [79] for a detailed illustration. In the classic demonstration, Ball A stops the moment it touches B, and B immediately starts to move, at the *same* speed A had been traveling. This visual display describes not only kinematic motions, but a causal interaction in which A "launches" B. Perception of this "launching effect" has a few notable properties that we enumerate below; see Ref. [119] for a more detailed review.



(a)

(b)

Figure 11: (a) An animation illustrates the intent, mood, and role of the agents [73]. The motion and interaction of four different pucks moving on a 2D plane are governed by latent physical properties and dynamic laws such as mass, friction, and global and pairwise forces. (b) Intuitive theory and counterfactual reasoning about the dynamics of the scene [74]. Schematic diagram of a collision event between two billiard balls, A and B, where the solid lines indicate the balls' actual movement paths and the dashed line indicates how Ball B would have moved if Ball A had not been present in the scene.

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. Tight control by spatial-temporal patterns of motion: By adding even a small temporal gap between the stop of A and the motion of B, perception of the launching effect will break down; instead, B's motion will be perceived as self-propelled.
3. Richness: Even the interaction of only two balls can support a variety of causal effects. For example, if B moves with a speed *faster* (vs. the same) than that of 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 after they make contact.

Recent cognitive science studies [134] provide still more striking evidence of how deeply human vision is rooted in causality, making the comparison between color and causality still more profound. In human vision science, "adaptation" is a phenomenon in which an observer adapts to stimuli after a period of sustained viewing, such that their perceptual response to those stimuli becomes weaker. In a particular type of adaptation, 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 adaptation has been taken as strong evidence of early visual processing of that stimuli. For example, it is well-known that the perception of color can induce retinotopic adaptation [135]. Strikingly, recent evidence revealed that retinotopic adaptation also takes place for the perception of causality. After prolonged viewing of the launching effect, subsequently viewed displays were judged more often as non-causal only if the displays were located within the same retinotopic coordinates. This means that physical causality is extracted during early visual processing. By using retinotopic adaptation as a tool, Kominsky and Scholl [136] 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 goes beyond placing labels on different causal events. One unique function of causality is the support of counterfactual reasoning. Observers recruit their counterfactual reasoning capacity to interpret visual events. In other words, interpretation is not based only on what is observed, but also on what would have happened but did not. In one study [137], participants judged whether one billiard ball caused another to go or prevented it from going through a gate. The participants' viewing patterns and judgments demonstrated that the 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 plays 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 remains much debate in the literature as to whether causal relationship understanding is necessary for high-level machine intelligence. However, learning causal concepts is of the utmost importance to agents that are expected to operate in observationally varying domains with common latent dynamics. To make this concrete, our environment on Earth adheres to relatively constant environmental dynamics, such as constant gravity. Perhaps more importantly, much of our world is *designed* by other humans and largely adheres to common causal concepts: Switches turn things off and on, knobs turn to open doors, and so forth. Even though objects in different settings appear different, their causal effect is constant because they all fit and cohere to a consistent causal design. Thus, for agents expected to work in varying but human-designed environments, the ability to learn generalizable and transferable causal understanding is crucial.

Recent successes of systems such as deep reinforcement learning (RL) showcase a broad range of applications [138, 139, 140, 141, 142], the vast majority of which do not learn explicit causal relationships. This results in a significant challenge for transfer learning under today's dominant machine learning paradigm [143, 144]. One approach to solving this challenge is to learn a causal encoding of the environment, because 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 an object's color, shape, or position).

In a study, Edmonds *et al.* [132] presented a complex hierarchical task that requires humans to reason about abstract causal structure. The work proposed a set of virtual "escape rooms," where agents must manipulate a series of levers to open a door; see an example in Fig. 12 [132]. Critically, this task is designed to force agents to form a causal structure by requiring agents to find *all* the ways to escape the room, rather than just one. The work used three- and four-lever rooms and two causal structures: Common Cause (CC) and Common Effect (CE). These causal structures encode different combinations into the rooms' locks.

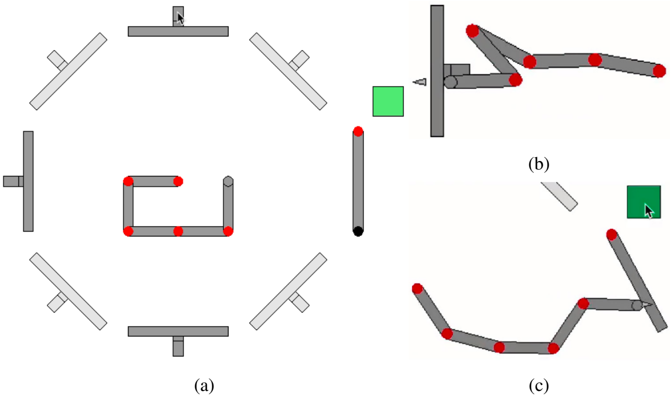
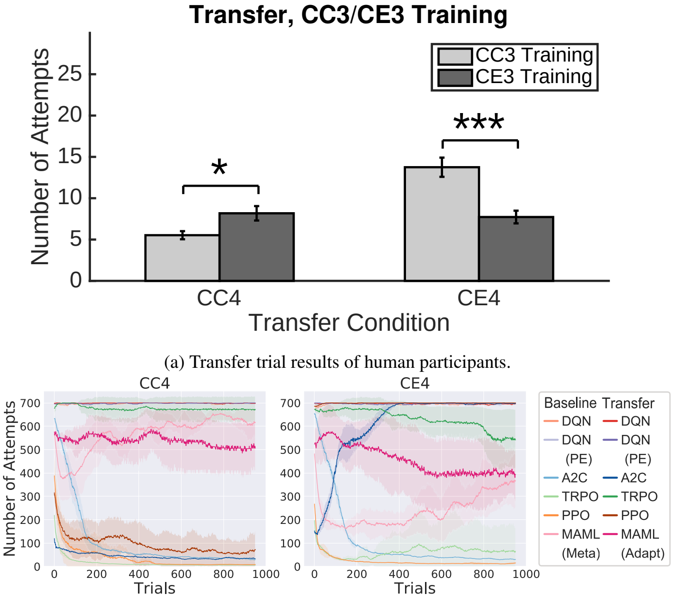


Figure 12: The OpenLock task presented in Ref. [132]. (a) Starting configuration of a three-lever trial. All levers are being pulled toward the robot arm, whose base is anchored to the center of the display. The arm interacts with levers by either *pushing* outward or *pulling* inward. This is achieved 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)-trained agents 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) Pushing a lever. (c) Opening the door by clicking the green button

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 the relationship between its perception of the new environment and the same latent causal structure on the fly. Finally, 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 its acquired knowledge to more complex circumstances.

This task setting is unique and challenging for two major reasons: (i) transferring agents between rooms tests whether or not agents form *abstract* representations of the environment; and (ii) transferring between three- and four-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 Fig. 13 [130, 145]. Humans approached optimal performance and showed positive transfer effects in rooms with an additional lever in both congruent and incongruent conditions. In contrast, recent deep RL methods failed to account for necessary causal abstraction, and showed a negative transfer effect. These results suggest that systems operating under



(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 [145]. Common Cause 4 (CC4) and Common Effect 4 (CE4) denote two transfer conditions used by Edmonds *et al.* [132]. (a) Average number of attempts human participants needed to find all unique solutions under four-lever Common Cause (CC4; left) and Common Effect (CE4; right) conditions, showing a positive causal transfer after learning. Light and dark gray 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 that the best-performing algorithms (Proximal Policy Optimization (PPO), Trust Region Policy Optimization (TRPO)) achieve success in 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. DQN: deep Q-network; DQN (PE): deep Q-network with prioritized experience replay; MAML: model-agnostic meta-learning.

current machine learning paradigms cannot learn a proper abstract encoding of the environment; that is, they do not learn an abstract causal encoding. Thus, we treat learning causal understanding from perception and interaction as one type of "dark matter" facing current AI systems, which should be explored further in future work.

3.3. *Causality in Statistical Learning*

Rubin [146] laid the foundation for causal analysis in statistical learning in his seminal paper, "Estimating causal effects of treatments in randomized and nonrandomized studies;" see also Ref. [147]. The formulation this work demonstrated is commonly called the Rubin causal model. The key concept in the Rubin causal model is potential outcomes. In the simplest scenario, where there are two treatments for each subject (*e.g.*, smoking or not smoking), the causal effect is defined as the difference between potential outcomes under the two treatments. The difficulty with causal inference is that, for each subject, we only observe the outcome under the one treatment that is actually assigned to the subject; the potential outcome, if the other treatment had been assigned to that subject, is missing. If the assignment of the treatment to each subject depends on the potential outcomes under the two treatments, a naive analysis comparing the observed average outcomes of the treatments that are actually assigned to the subjects will result in misleading conclusions. A common manifestation of this problem is the latent variables that influence both the treatment assignment and the potential outcomes (*e.g.*, a genetic factor influencing both one's tendency to smoke and one's health). A large body of research has been developed to solve this problem. A very prominent example is the propensity score [148], which is the conditional probability of assigning one treatment to a subject given the background variables of the subject. Valid causal inference is possible by comparing subjects with similar propensity scores.

Causality was further developed in Pearl's probabilistic graphical model (*i.e.*, causal Bayesian networks (CBNs)) [149]. CBNs enabled economists and epidemiologists to make inferences for quantities that cannot be intervened upon in the real world. Under this framework, an expert modeler typically 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 causal relationships from observational data [150]. These efforts tried to determine under what circumstances the structure (presence and direction of an edge between two variables in CBN) could be determined from purely observational data [150, 151, 152].

This framework is a powerful tool in fields where real-world interventions are difficult (if not impossible)—such as economics and epidemiology—but lacks many properties necessary for humanlike AI. First, despite attempts to learn causal structure from observational data, most structure learning approaches cannot typically succeed beyond identifying a Markov equivalence class of possible structures [152]; therefore, structure learning remains an unsolved problem. Recent work has attempted to tackle this limitation by introducing *active intervention* that enables agents to explore possible directions of undirected causal edges [153, 154]. 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 relationships must be considered. Humanlike AI should have the ability to constrain the space of possible relationships to what is heuristically "reasonable" given the agent's understanding of the world, while acknowledging that such a learning process may not result in the ground-truth causal model. That is, we suggest that for building humanlike AI, learners should relax the formalism imposed by CBNs to accommodate significantly more variables without disregarding explicit causal structure (as is currently done by 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 world model is updated with new confirming or contradictory evidence.

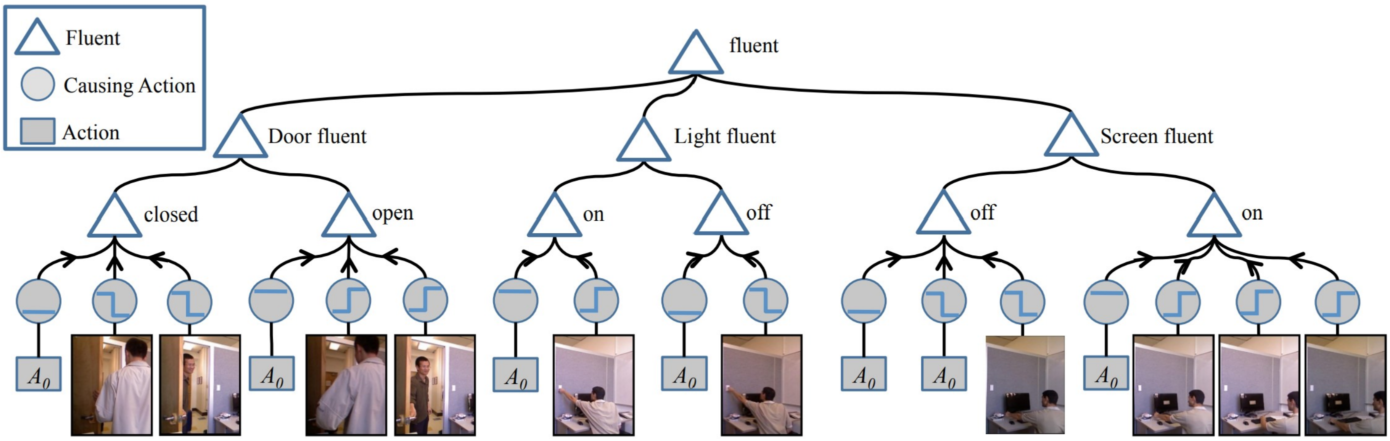


Figure 14: An example of perceptual causality in computer vision [155], with a causal and-or graph for door status, light status, and screen status. Action A_0 represents non-action (a lack of state-changing agent action). Non-action is also used to explain the change of the monitor status to off when the screensaver activates. Arrows point from causes to effects, and undirected lines show deterministic definition.

3.4. Causality in Computer Vision

The classical and scientific clinical setting for learning causality is Fisher's randomized controlled experiments [156]. 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, it provides a stark contrast to the human ability to perceive causal relationships from observations alone [119, 127, 128]. These works suggest that human causal perception is less rigorous than formal science but still maintains effectiveness in learning and understanding of daily events.

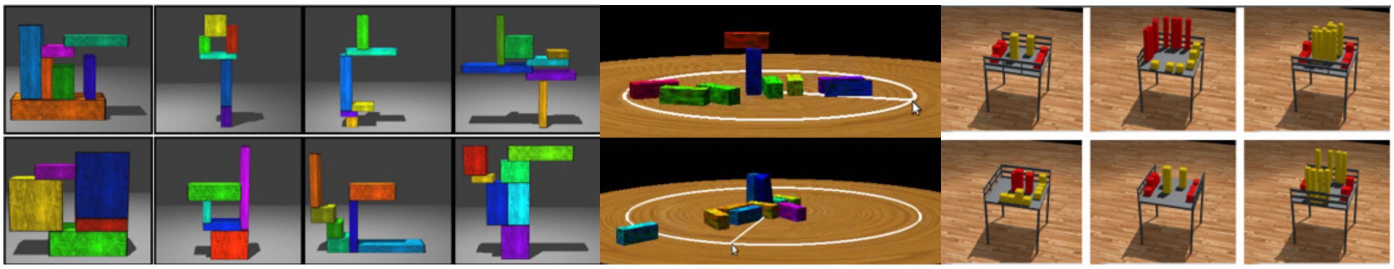
Accordingly, computer vision and AI approaches should focus on how humans perceive causal relationships from observational data. Fire and Zhu [157, 155] proposed a method to learn "dark" causal relationships from image and video inputs, as illustrated in Fig. 14 [157]; in this study, systems learn how the status of a door, light, and screen relate to human actions. Their method achieves this iteratively by asking the same question at different intervals: *given the observed videos and the current causal model, what causal relationship should be added to the model to best match the observed statistics describing the causal events?* To answer this question, the method utilizes the information projection framework [158], maximizing the amount of information gain after adding a causal relation, 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, and so forth. Under the information projection framework, the top-scoring causal relationships consistently matched what humans perceived to be a cause of action in the scene, while low-scoring causal relations matched what humans perceived to *not* be a cause of action in the scene. These results indicate that 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 are not guaranteed to uncover the complete and true causal structure, perceptual causality provides a mechanism to achieve humanlike learning from observational data.

Causality is crucial for humans' understanding and reasoning about videos, such as tracking humans that are interacting with objects whose visibility might vary over time. Xu *et al.* [159] used a Causal And-Or Graph (C-AOG) model to tackle this kind of "visibility fluent reasoning" problem. They consider the visibility status of an object as a fluent variable, whose change is mostly attributed to its interaction with its surroundings, such as crossing behind another object, entering a building, or getting into a vehicle. The proposed C-AOG can represent the cause-effect relationship between an object's activities and its visibility fluent; based on this, the researchers developed a probabilistic graphical model to jointly reason about the visibility fluent change and track humans. Experimental results demonstrate that with causal reasoning, they can recover and describe complete trajectories of humans interacting frequently in complicated scenarios. Xiong *et al.* [160] also defined causality as a fluent change due to relevant action, and used a C-AOG to describe the causal understanding demonstrated by robots that successfully folded clothes after observing humans doing the same.

4. Intuitive Physics: Cues of the Physical World

Perceiving causality, and using this perception to interact with an environment, requires a commonsense understanding of how the world operates at a physical level. Physical understanding does not necessarily require us to precisely or explicitly invoke Newton's laws of mechanics; instead, we rely on intuition, built up through 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* [161]. The field of intuitive physics has been explored for several decades in cognitive science and was recently reinvigorated by new techniques linked to AI.



(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 about physics, stability, and support relationships presented in Ref. [70]: 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 judgment of physics can be viewed as a form of probabilistic inference over the principles of Newtonian mechanics.

Surprisingly, humans develop physical intuition at an early age [80], 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 makes visual task completion an important goal for future machine vision and AI systems. We begin this section with a short review of intuitive physics in human cognition, followed by a review of recent developments in computer vision and AI that use 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 in which 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 point in time [162, 161, 163]. However, humans' intuitive understanding of physics was shown later to be much more accurate, rich, and sophisticated than previously expected once *dynamics* and proper *context* were provided [164, 165, 166, 167, 168].

These later findings are fundamentally different from prior work that systematically investigated the development of infants' physical knowledge [169, 170] in the 1950s. The reason for such a difference in findings is that the earlier research included not only tasks of merely reasoning about physical knowledge, but also other tasks [171, 172]. To address such difficulties, researchers have developed alternative experimental approaches [173, 93, 174, 175] to study the development of infants' physical knowledge. The most widely used approach is the violation-of-expectation method, in which infants see two test events: an expected event, consistent with the expectation shown, and an unexpected event, violating the expectation. A series of these kinds of studies have provided strong evidence that humans—even young infants—possess expectations about a variety of physical events [176, 177].

In a single glance, humans can perceive whether a stack of dishes will topple, whether a branch will support a child's weight, whether a tool can be lifted, and whether an object can be caught or dodged. In these complex and dynamic events, the ability to perceive, predict, and therefore appropriately interact with objects in the physical world relies on 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, Achinstein [178] argued that the brain builds mental models to support inference through mental simulations, analogous to how engineers use simulations for the 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 [179] suggesting that systematic parietal and frontal regions are engaged when humans perform physical inferences even when simply viewing physically rich scenes. These findings suggest that these brain regions use a generalized mental engine for intuitive physical inference—that is, the brain's "physics engine." These brain regions are much more active when making physical inferences relative to when making inferences about *nonphysical* but otherwise highly similar scenes and tasks. Importantly, these regions are not exclusively engaged in physical inference, but are also overlapped with the brain regions involved in action planning and tool use. This indicates a very intimate relationship between the cognitive and neural mechanisms for understanding intuitive physics, and the mechanisms for preparing appropriate actions. This, in turn, is a critical component linking perception to action.

To construct humanlike commonsense knowledge, a computational model for intuitive physics that can support the performance of *any* task that involves physics, not just one narrow task, must be explicitly represented in an agent's environmental understanding. This requirement stands against the recent "end-to-end" paradigm in AI, in which neural networks directly map an input image to an output action for a specific task, leaving an implicit internal task representation "baked" into the network's weights.

Recent breakthroughs in cognitive science provide solid evidence supporting the existence of an intuitive physics model in human scene understanding. This evidence suggests that humans perform physical inferences by running probabilistic

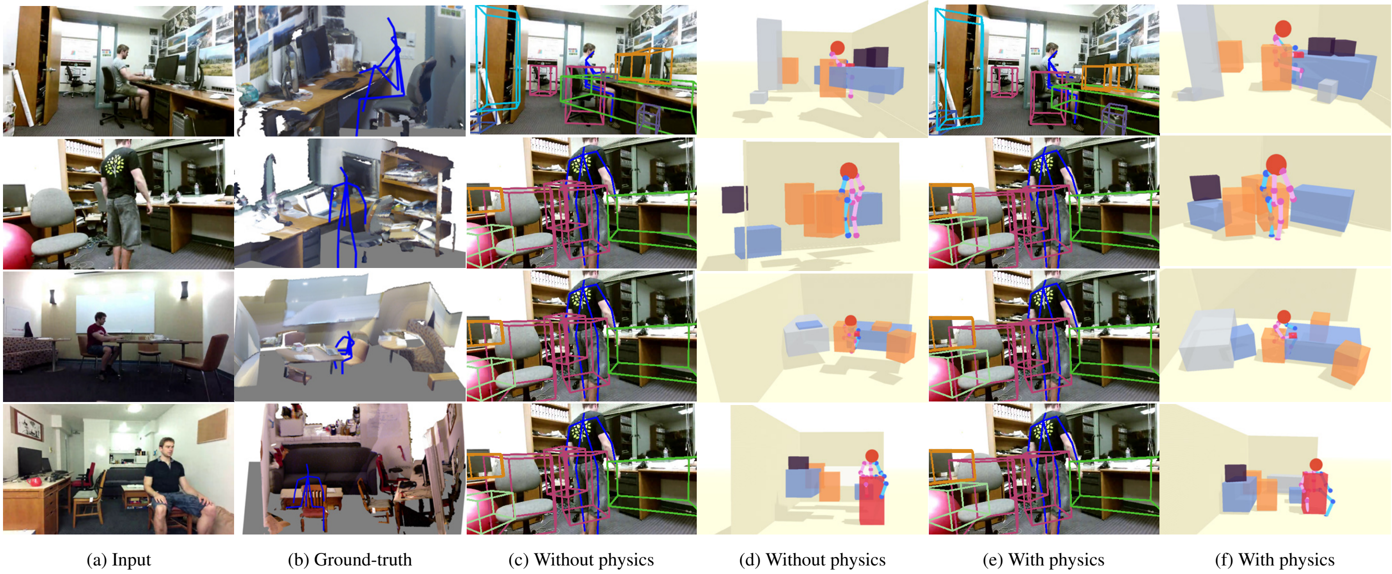


Figure 16: Scene parsing and reconstruction by integrating physics and human-object interactions. (a) Input image; (b) ground truth; (c and d) without incorporating physics, the objects might appear to float in the air, resulting in an incorrect parsing; (e and f) after incorporating physics, the parsed 3D scene appears physically stable. The system has been able to perceive the "dark" physical stability in which objects must rest on one another to be stable. Reproduced from Ref. [37] with permission of IEEE, © 2019.

simulations in a mental physics engine akin to the 3D physics engines used in video games [180]; see Fig. 15 [70]. Human intuitive physics can be modeled as an approximated physical engine with a Bayesian probabilistic model [70], possessing the following distinguishing properties: (i) Physical judgment is achieved by running a coarse and rough forward physical simulation; and (ii) the simulation is stochastic, which is different from the deterministic and precise physics engine developed in computer graphics. For example, in the tower stability task presented in Ref. [70], there is uncertainty about the exact physical attributes of the blocks; they fall into a probabilistic distribution. For every simulation, the model first samples the blocks' attributes, then generates predicted states by recursively applying elementary physical rules over short-time intervals. This process creates a distribution of simulated results. The stability of a tower is then represented in the results as the probability of the tower not falling. Due to its stochastic nature, this model will judge a tower as stable only when it can tolerate small jitters or other disturbances to its components. This single model fits data from five distinct psychophysical tasks, captures several illusions and biases, and explains core aspects of mental models and commonsense reasoning that are instrumental to how humans understand their everyday world.

More recent studies have demonstrated that intuitive physical cognition is not limited to the understanding of rigid bodies, but also expands to the perception and simulation of the physical properties of liquids [181, 182] and sand [183]. In these studies, the experiments demonstrate that humans do not rely on simple qualitative heuristics to reason about fluid or granular dynamics; instead, they rely on perceived physical variables to make quantitative judgments. Such results provide converging evidence supporting the idea of mental simulation in physical reasoning. For a more in-depth review of intuitive physics in psychology, see Ref. [184].

4.2. *Physics-based Reasoning in Computer Vision*

Classic computer vision studies focus on reasoning about appearance and geometry—the highly visible, pixel-represented aspects of images. Statistical modeling [185] 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 [186]." Marr conjectured that the perception of a 2D image is an *explicit* multiphase information process [1], involving (i) an early vision system for perceiving [187, 188] and textons [189, 190] to form a primal sketch [191, 192]; (ii) a mid-level vision system to form 2.1D [193, 194, 195] and 2.5D [196] sketches; and (iii) a high-level vision system in charge of full 3D scene formation [197, 198, 199]. In particular, Marr highlighted the importance of different levels of organization and the internal representation [200].

Alternatively, perceptual organization [201, 202] and Gestalt laws [203, 204, 205, 206, 207, 208, 209, 210] aim to resolve the 3D reconstruction problem from a single RGB image without considering depth. Instead, they use priors—groupings and structural cues [211, 212] that are likely to be invariant over wide ranges of viewpoints [213]—resulting in feature-based approaches [214, 88].

However, both appearance [215] and geometric [29] approaches have well-known difficulties resolving ambiguities. In addressing this challenge, modern computer vision systems have started to account for "dark" aspects of images by incorporating physics; as a result, they have demonstrated dramatic improvements over prior works. In certain cases, ambiguities have been shown to be extremely difficult to resolve through current state-of-the-art data-driven classification methods, indicating the significance of "dark" physical cues and signals in our ability to correctly perceive and operate within our daily

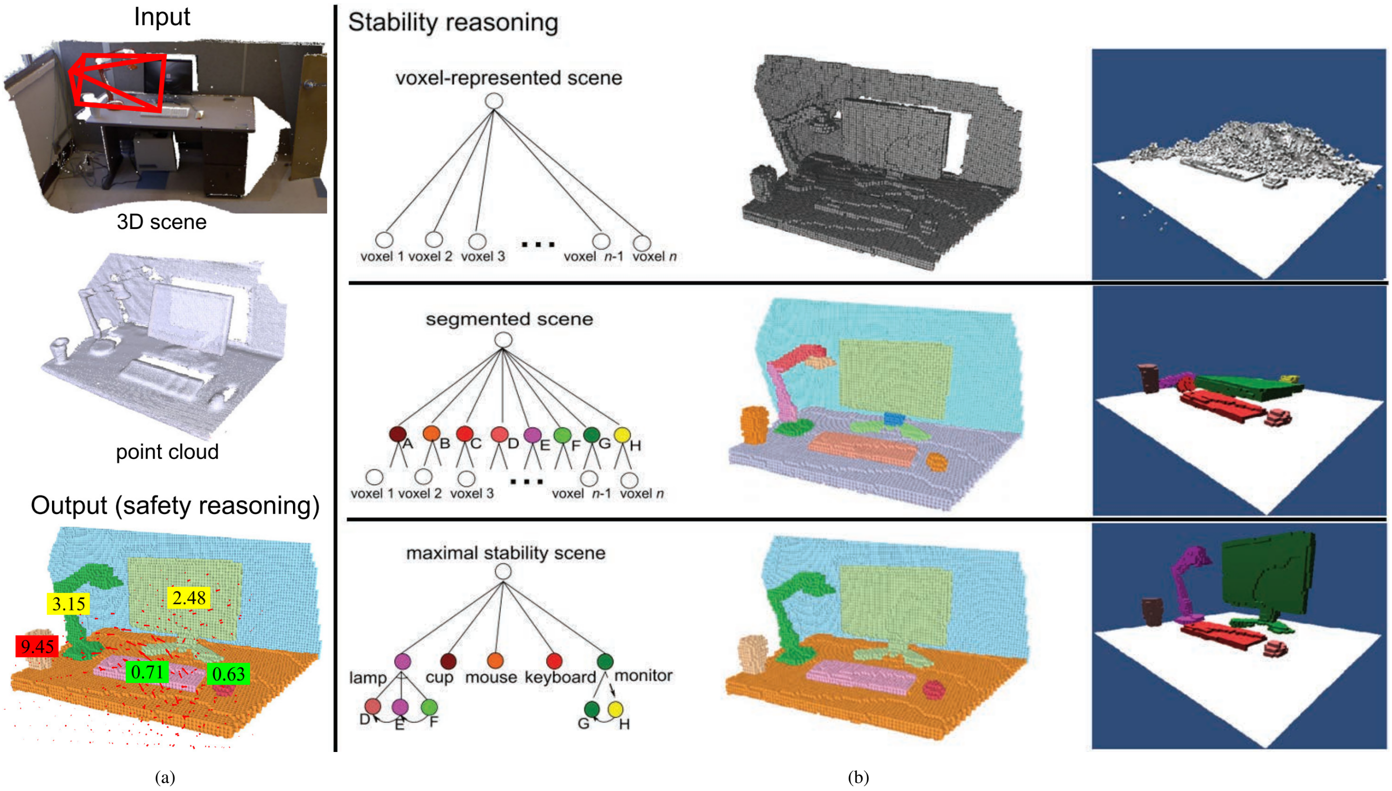


Figure 17: An example explicitly exploiting safety and stability in a 3D scene-understanding task. Good performance in this task means that the system can understand the "dark" aspects of the image, which include how likely each object is to fall, and where the likely cause of falling will come from. (a) Input: reconstructed 3D scene. Output: parsed and segmented 3D scene comprised of stable objects. The numbers are "unsafety" scores for each object with respect to the disturbance field (represented by red arrows). (b) Scene-parsing graphs corresponding to three bottom-up processes: voxel-based representation (top), geometric pre-process, including segmentation and volumetric completion (middle), and stability optimization (bottom). Reproduced from Ref. [98] with permission of Springer Science+Business Media New York, © 2015.

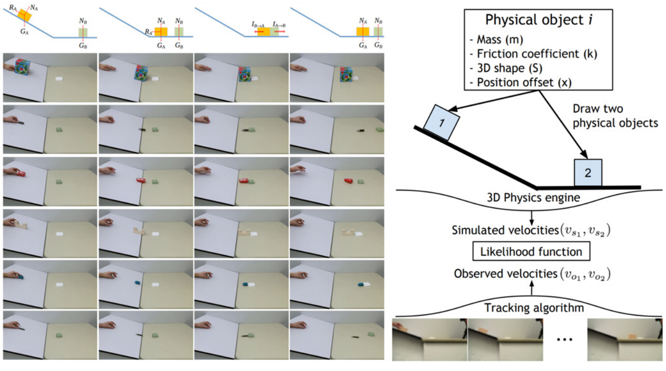
environments; see examples in Fig. 16 [37], where systems perceive which objects must rest on each other in order to be stable in a typical office space.

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

1. Stability and safety in scene understanding. As demonstrated in Ref. [98], this line of work is mainly based on a simple but crucial observation in human-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 physically plausible interpretation in scene understanding.
2. Physical relationships in 3D scenes. Humans excel in reasoning about the physical relationships in a 3D scene, such as which objects support, attach, or hang from one another. As shown in Ref. [36], those relationships represent a deeper understanding of 3D scenes beyond observable pixels that could benefit a wide range of applications in robotics, virtual reality (VR), and augmented reality (AR).

The idea of incorporating physics to address vision problems can be traced back to Helmholtz and his argument for the "unconscious inference" of probable causes of sensory input as part of the formation of visual impressions [216]. The very first such formal solution in computer vision dates back to Roberts' solutions for the parsing and reconstruction of a 3D block world in 1963 [217]. This work inspired later researchers to realize the importance of both the violation of physical laws for scene understanding [218] and stability in generic robot manipulation tasks [219, 220].

Integrating physics into scene parsing and reconstruction was revisited in the 2010s, bringing it into modern computer vision systems and methods. From a single RGB image, Gupta *et al.* proposed a qualitative physical representation for indoor [31, 101] and outdoor [221] scenes, where an algorithm infers the volumetric shapes of objects and relationships (such as occlusion and support) in describing 3D structure and mechanical configurations. In the next few years, other work [222, 223, 224, 225, 226, 109, 32, 227, 228, 34] also integrated the inference of physical relationships for various scene understanding tasks. In the past two years, Liu *et al.* [35] inferred physical relationships in joint semantic segmentation and 3D reconstruction of outdoor scenes. Huang *et al.* [36] modeled support relationships as edges in a human-centric scene graphical model, inferred the relationships by minimizing supporting energies among objects and the room layout, and enforced physical stability and plausibility by penalizing the intersections among reconstructed 3D objects and room layout [100, 37].


Physical object i
- Mass (m)
- Friction coefficient (k)
- 3D shape (S)
- Position offset (x)
Draw two physical objects
1
2
3D Physics engine
Simulated velocities(v_{s_1}, v_{s_2})
Likelihood function
Observed velocities(v_{o_1}, v_{o_2})
Tracking algorithm
...

(a) Snapshots of datasets

(b) Galileo model

Figure 18: Inferring the dynamics of the scenes. (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. Reproduced from Ref. [230] with permission of Neural Information Processing Systems Foundation, Inc., © 2015

The aforementioned recent work mostly adopts simple physics cues; that is, very limited (if any) physics-based simulation is applied. The first recent work that utilized an actual physics simulator in modern computer vision methods was proposed by Zheng *et al.* in 2013 [97, 81, 98]. As shown in Fig. 17 [98], the proposed method first groups potentially unstable objects with stable ones by optimizing for stability in the scene prior. Then, it assigns an "unsafety" prediction score to each potentially unstable object by inferring hidden potential triggers of instability (the disturbance field). The result is a physically plausible scene interpretation (voxel segmentation). This line of work has been further explored by Du *et al.* [229] by integrating an end-to-end trainable network and synthetic data.

Going beyond stability and support relationships, Wu *et al.* [230] integrated physics engines with deep learning to predict the future dynamic evolution of static scenes. Specifically, a generative model named Galileo was proposed for physical scene understanding using real-world videos and images. As shown in Fig. 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. Wu *et al.* [231] further explored directly mapping visual inputs to physical properties, inverting a part of the generative process using deep learning. Object-centered physical properties such as mass, density, and the 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 (*i.e.*, shapes, colors, and sizes), the proposed unsupervised representation learning model, which explicitly encodes basic physical laws into the structure, can learn the physical properties of objects from videos.

Integrating physics and predicting future dynamics opens up quite a few interesting doors in computer vision. For example, given a human motion or task demonstration presented as a RGB-D image sequence, *et al.* [232] built a system that calculated various physical concepts from just a single example of tool use (Fig. 19), enabling it to reason about the essential physical concepts of the task (*e.g.*, the force required to crack nuts). As the fidelity and complexity of the simulation increased, Zhu *et al.* [233] were able to infer the forces impacting a seated human body, using a finite element method (FEM) to generate a mesh estimating the force on various body parts; Fig. 35d.

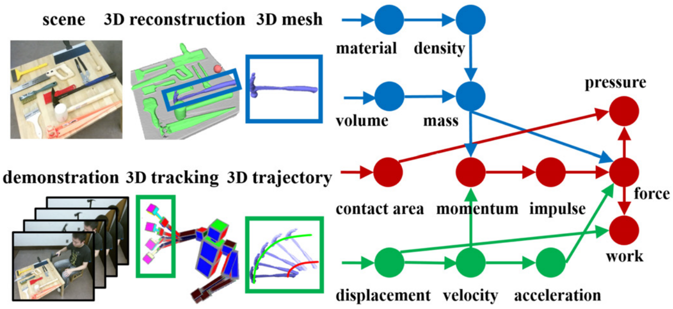


Figure 19: Thirteen physical concepts involved in tool use and their compositional relationships. By parsing a human demonstration, the physical concepts of material, volume, concept area, and displacement are estimated from 3D meshes of tool attributes (blue), trajectories of tool use (green), or both together (red). Higher level physical concepts can be further derived recursively. Reproduced from Ref. [232] with permission of the authors, © 2015.

Physics-based reasoning can not only be applied to scene understanding tasks, as above, but have also been applied to pose and hand recognition and analysis tasks. For example, Brubaker *et al.* [234, 235, 236] estimated the force of contacts and the torques of internal joints of human actions using a mass-spring system. Pham *et al.* [237] further attempted to infer the forces of hand movements during human-object manipulation. In computer graphics, soft-body simulations based on video observation have been used to jointly track human hands and calculate the force of contacts [238, 239]. Altogether, the laws of physics and how they relate to and among objects in a scene are critical "dark" matter for an intelligent agent to perceive and understand; some of the most promising computer vision methods outlined above have understood and incorporated this insight.

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

Perception of an environment inevitably leads to a course of action [240, 241]; Gibson argued that clues indicating opportunities for action in a nearby environment are perceived in a *direct*, *immediate* way with no sensory processing. This is particularly true for human-made objects and environments, as "an object is first identified as having important functional relations" and "perceptual analysis is derived of the functional concept" [242]; for example, switches are clearly for flipping, buttons for pushing, knobs for turning, hooks for hanging, caps for rotating, handles for pulling, and so forth. This idea is the core of affordance theory [243], which is based on Gestalt theory and has had a significant influence on how we consider visual perception and scene understanding.

Functional understanding of objects and scenes is rooted in identifying possible tasks that can be performed with an object [244]. This is deeply related to the perception of causality, as covered in Section 3; to understand how an object can be used, an agent must understand what change of state will result if an object is interacted with in any way. While affordances depend directly on the actor, functionality is a permanent property of an object independent of the characteristics of the user; see an illustration of this distinction in Fig. 21. These two interweaving concepts are more invariant for object and scene understanding than their geometric and appearance aspects. Specifically, we argue that:

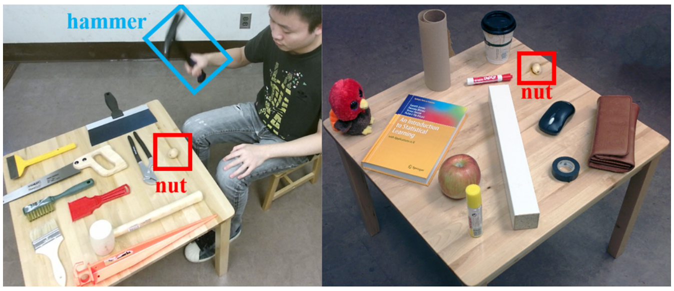
1. Objects, especially human-made ones, are defined by their functions, or by the actions they are associated with;
2. Scenes, especially human-made ones, are defined by the actions than can be performed within them.

Functionality and affordance are interdisciplinary topics and have been reviewed from different perspectives in the literature (*e.g.*, Ref. [245]). In this section, we emphasize the importance of incorporating functionality and affordance in the field of computer vision and AI by starting with a case study of tool use in animal cognition. A review of functionality and affordance in computer vision follows, from both the object level and scene level. At the end, we review some recent literature in robotic manipulation that focuses on identifying the functionality and affordance of objects, which complements previous reviews of data-driven approaches [246] and affordance tasks [247].

5.1. *Revelation from Tool Use in Animal Cognition*

The ability to use an object as a tool to alter another object and accomplish a task has traditionally been regarded as an indicator of intelligence and complex cognition, separating humans from other animals [248, 249]. Researchers commonly viewed tool use as the hallmark of human intelligence [250] until relatively recently, when Dr. Jane Goodall observed wild chimpanzees manufacturing and using tools with regularity [251, 252, 253]. Further studies have since reported on tool use by other species in addition to chimpanzees. For example, Santos *et al*. [254] trained two species of monkeys to choose between two canes to reach food under a variety of conditions involving different types of physical concepts (*e.g.*, materials, connectivity, and gravity). Hunt [255] and Weir *et al*. [256] reported that New Caledonian crows can bend a piece of straight wire into a hook and use it to lift a bucket containing food from a vertical pipe. More recent studies also found that New Caledonian crows behave optimistically after using tools [257]. Effort cannot explain their optimism; instead, they appear to enjoy or be intrinsically motivated by tool use.

These discoveries suggest that some animals have the capability (and possibly the intrinsic motivation) to reason about the functional properties of tools. They can infer and analyze physical concepts and causal relationships of tools to approach a novel task using domain-general cognitive mechanisms, despite huge variety in their visual appearance and geometric features. Tool use is of particular interest and poses two major challenges in comparative cognition [258], which further challenges the reasoning ability of computer vision and AI systems.



(a) Learning

(b) Inference

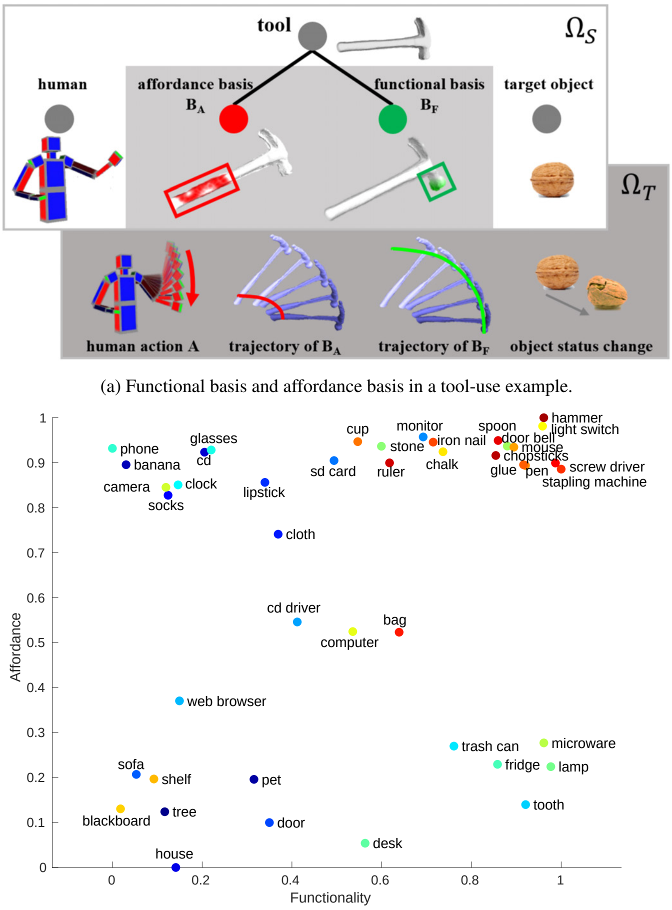
Figure 20: Finding the right tools in novel situations. (a) In a learning phase, a rational human charged with cracking a nut is observed examining a hammer and other tools; (b) in an inference phase, the algorithm is asked to pick the best object on the table (*i.e.*, the wooden leg) for the same task. This generalization entails reasoning about functionality, physics, and causal relationships among objects, actions, and overall tasks. Reproduced from Ref. [232] with permission of the authors, © 2015.

First, why can some species devise innovative solutions, while others facing the same situation cannot? Look at the example in Fig. 20 [232]: by observing only a single demonstration of a person achieving the complex task of cracking a nut, we humans can effortlessly reason about which of the potential candidates from a new set of random and very different objects is best capable of helping us complete the same task. Reasoning across such large intraclass variance is extremely difficult to capture and describe for modern computer vision and AI systems. Without a consistent visual pattern, properly identifying tools for a given task is a long-tail visual recognition problem. Moreover, the very same object can serve multiple functions depending on task context and requirements. Such an object is no longer defined by its conventional name (*i.e.*, a hammer); instead, it is defined by its functionality.

Second, how can this functional reasoning capability emerge if one does not possess it innately? New Caledonian crows are well-known for their propensity and dexterity at making and using tools; meanwhile, although a crow's distant cousin, the rook, is able to reason and use tools in a lab setting, even *they* do not use tools in the wild [259]. These findings suggest that the ability to represent tools may be more of a domain-general cognitive capacity based on functional reasoning 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 [243]



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

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

Figure 21: (a) The task-oriented representation of a hammer and its use in cracking a nut in a joint spatiotemporal space. In this example, an object is decomposed into a functional basis and an affordance basis for a given task. (b) The likelihood of a common object being used as a tool based on its functionality and affordance. The warmer the color, the higher the probability. The functionality score is the average response to the question "Can it be used to change the status of another object?", and the affordance score is the average response to "Can it be manipulated by hand?"

The idea to incorporate functionality and affordance into computer vision and AI can be dated back to the second International Joint Conference on Artificial Intelligence (IJCAI) in 1971, where Freeman and Newell [260] argued that available structures should be described in terms of functions provided and functions performed. The concept of affordance was later coined by Gibson [243]. Based on the classic geometry-based "arch-learning" program [261], Winston *et al.* [262] discussed the use of function-based descriptions of object categories. They pointed out that it is possible to use a single functional description to represent all possible cups, despite there being an infinite number of individual physical descriptions of cups or many other objects. In their "mechanic's mate" system [263], Connell and Brady [264] proposed semantic net descriptions based on 2D shapes together with a generalized structural description. "Chair" and "tool," exemplary categories researchers used for studies in functionality and affordance, were first systematically discussed alongside a computational method by Ho [265] and DiManzo *et al.* [266], respectively. Inspired by the functional aspect of the "chair" category in Minsky's book [267], 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 and Bowyer [268]. 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 topics.

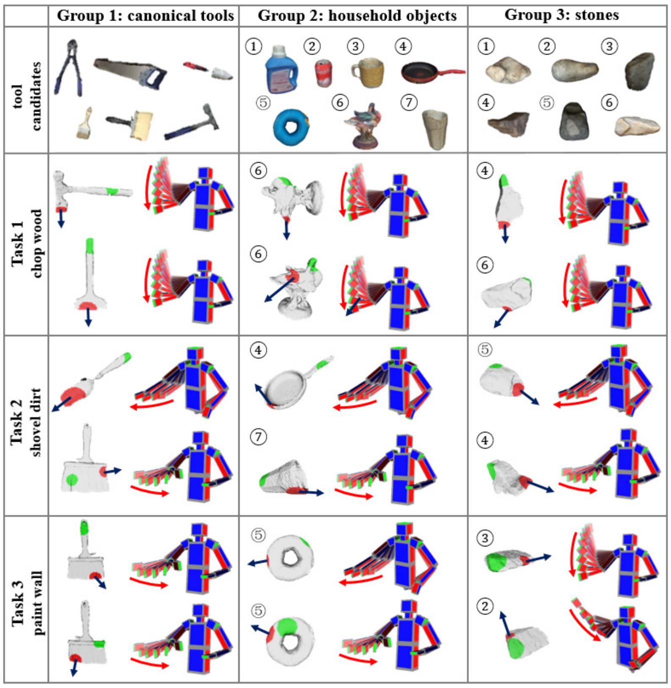


Figure 22: Given the three tasks of chopping wood, shoveling dirt, and painting a wall, an algorithm proposed by Zhu *et al.* [232] picks and ranks objects within groups in terms of which object in each group is the best fit for task performance: conventional tools, household objects, and stones. Second, the algorithm outputs the imagined use of each tool, providing an affordance basis (the green spot indicating where the tool would be grasped by hand), a functional basis (the red area indicating the part of the tool that would make contact with the object), and the imagined sequence of poses constituting the movement of the action itself. Reproduced from Ref. [232] with permission of the authors, © 2015.

"Tool" is of particular interest in computer vision and robotics, partly due to its nature as an object for changing *other* objects' status. Motivated by the studies of tool use in animal cognition, Zhu *et al.* [232] cast the tool understanding problem as a *task-oriented* object-recognition problem, the core of which is understanding an object's underlying functions, physics, and causality. As shown in Fig. 22 [232], a tool is a physical object (*e.g.*, a hammer or a shovel) that is used through action to achieve a task. From this new perspective, any object can be viewed as a hammer or a shovel. This generative representation allows computer vision and AI algorithms to reason about the underlying mechanisms of various tasks and generalize object recognition across novel functions and situations. This perspective goes beyond memorizing examples for each object category, which tends to prevail among traditional appearance-based approaches in the literature. Combining both

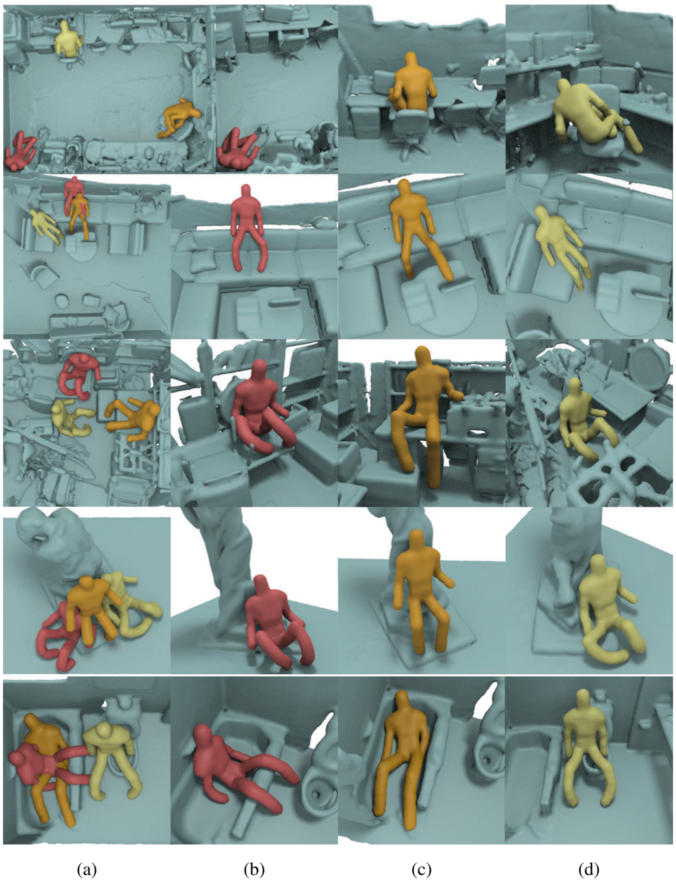


Figure 23: (a) Top three poses in various scenes for affordance (sitting) recognition. The zoom-in shows views of the (b) best, (c) second-best, and (d) third-best choice of sitting poses. The top two rows are canonical scenarios, the middle row is a cluttered scenario, and the bottom two rows are novel scenarios that demonstrated significant generalization and transfer capability. Reproduced from Ref. [233] with permission of the authors, © 2016.

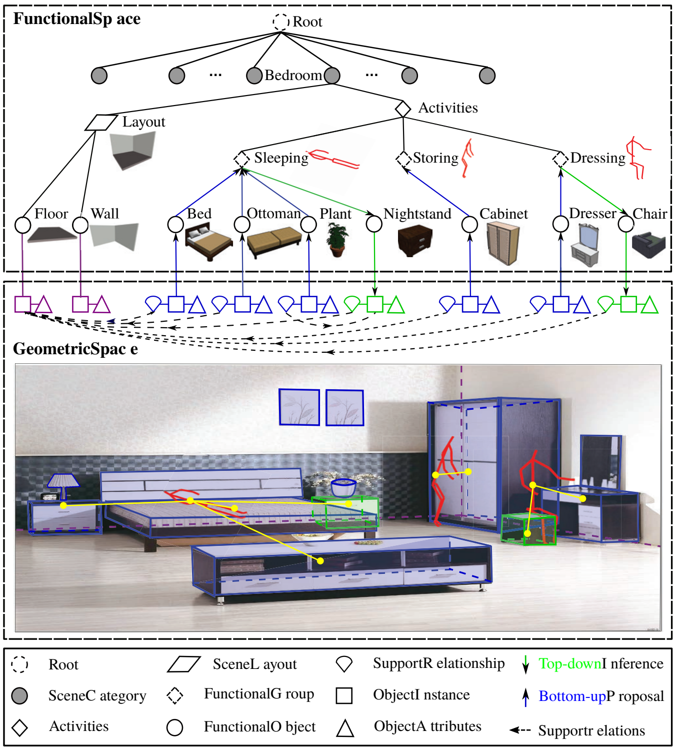


Figure 24: Task-centered representation of an indoor scene. The functional space exhibits a hierarchical structure, and the geometric space encodes the spatial entities with contextual relationships. The objects are grouped by their hidden activity, *i.e.*, by latent human context or action. Reproduced from Ref. [36] with permission of the authors, © 2018.

physical and geometric aspects, Liu *et al.* [269] took the decomposition of physical primitives for tool recognition and tower stability further.

"Container" is ubiquitous in daily life and is considered a half-tool [270]. The study of containers can be traced back to a series of studies by Inhelder and Piaget in 1958 [271], in which they showed six-year-old children could still be confused by the complex phenomenon of pouring liquid into containers. Container and containment relationships are of particular interest in AI, computer vision, and psychology due to the fact that it is one of the earliest spatial relationships to be learned, preceding other common ones *e.g.*, occlusions [272] and support relationships [273]). As early as two and a half months old, infants can already understand containers and containment [274, 275, 276]. In the AI community, researchers have been adopting commonsense reasoning [277, 278, 279] and qualitative representation [280, 281] for reasoning about container and containment relationships, mostly focusing on ontology, topology, first-order logic, and knowledge base.

More recently, physical cues and signals have been demonstrated to strongly facilitate reasoning about functionality and affordance in container and containment relationships. For example, Liang *et al.* [282] demonstrated that a physics-based simulation is robust and transferable for identifying containers in response to three questions: "What is a container?", "Will an object contain another?", and "How many objects will a container hold?" Liang's approach performed better than approaches using features extracted from appearance and geometry for the same problem. This line of research aligns with the recent findings of intuitive physics in psychology [70, 165, 181, 182, 183, 184], and enabled a few interesting new directions and applications in computer vision, including reasoning about liquid transfer [283, 284], container and containment relationships [285], and object tracking by utilizing containment constraints [286].

"Chair" is an exemplar class for affordance; the latest studies on object affordance include reasoning about both geometry and function, thereby achieving better generalizations for unseen instances than conventional, appearance-based, or geometry-based machine learning approaches. In particular, Grabner *et al.* [108] designed an "affordance detector" for chairs by fitting typical human sitting poses onto 3D objects. Going beyond visible geometric compatibility, through physics-based simulation, Zhu *et al.* [233] inferred the forces/pressures applied to various body parts while sitting on different chairs; see Fig. 23 [233] for more information. Their system is able to "feel," in numerical terms, discomfort when the forces/pressures on body parts exceed certain comfort intervals.

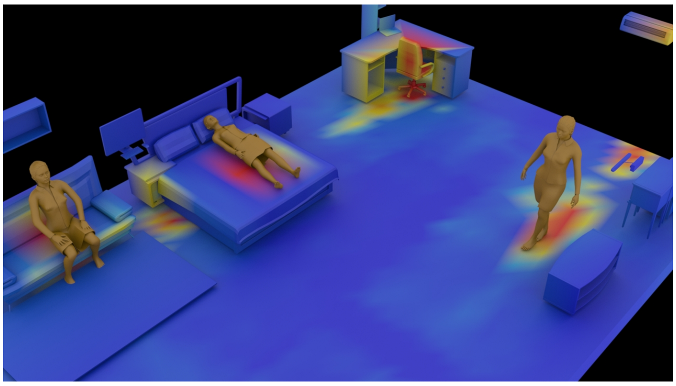
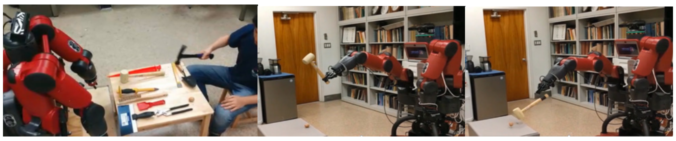


Figure 25: An example of a synthesized human-centric indoor scene (a bedroom) with an affordance heat map generated by Refs. [99, 288]. The joint sampling of the scene was achieved by alternatively sampling humans and objects according to a joint probability distribution.

"Human" context has proven to be a critical component in modeling the constraints on possible usage of objects in a scene. In approaching this kind of problem, all methods imagine different potential human positioning relative to objects to help parse and understand the visible elements of the scene. The fundamental reason for this approach is that human-made scenes are functional spaces that serve human activities, whose objects exist primarily to assist human actions [243]. Working at the object level, Jiang *et al.* proposed methods that use human context to learn object arrangement [287] and object labeling [110]. At the scene level, Zhao and Zhu [34] modeled functionality in 3D scenes through the compositional and contextual relationships among objects within them. To further explore the hidden human context pervading 3D scenes, Huang *et al.* [36] proposed a stochastic method to parse and reconstruct scenes with a holistic scene grammar (HSG). HSG describes a functional, task-centered representation of scenes. As shown in Fig. 24 [36], the descriptor was composed of functional scene categories, task-centered activity groups, and individual objects. In a reversal of the process of parsing scenes using human context, scene functionality could also be used to synthesize *new* scenes with humanlike object arrangements: Qi *et al.* [99] and Jiang *et al.* [288] proposed using human-centric representations to synthesize 3D scenes with a simulation engine. As illustrated in Fig. 25 [99, 288], they integrated human activities with functional grouping/support relationships to build natural and fitting activity spaces.

5.3. Mirroring: Causal-equivalent Functionality & Affordance

It is difficult to evaluate a computer vision or AI system's facility at reasoning with functionality and affordance; unlike with causality and physics, not all systems will see functionality and affordance in the same way. Indeed, humans and robots have different morphology; therefore, the same object or environment does not necessarily introduce the same functionality and affordance to both robots and humans. For example, a human with five fingers can firmly grasp a hammer that a robot gripper with the typical two or three fingers might struggle to wield, as shown in Fig. 26. In these cases, a system must reason about the underlying mechanisms of affordance, rather than simply mimicking the motions of a human demonstration. This common problem is known as the "correspondence problem" [289] in learning from demonstration (LfD); more details have been provided in two previous surveys [290, 291].



(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.

Currently, the majority of work in LfD uses a one-to-one mapping between human demonstration and robot execution, restricting the LfD to mimicking the human's low-level motor controls and replicating a nearly identical procedure. Consequently, the "correspondence problem" is insufficiently addressed, and the acquired skills are difficult to adapt to new robots or new situations; thus, more robust solutions are necessary. To tackle these problems, we argue that the robot must obtain deeper understanding in functional and causal understanding of the manipulation, which demands more explicit modeling of knowledge about physical objects and forces. The key to imitating manipulation is using functionality and affordance to create causal-equivalent manipulation; in other words, replicating task execution by reasoning about contact forces, instead of simply repeating the precise trajectory of motion.

However, measuring human manipulation forces is difficult due to the lack of accurate instruments; there are constraints imposed on devices aimed at measuring natural hand motions. For example, a vision-based force-sensing method [237] often cannot handle self-occlusions and occlusions caused during manipulations. Other force-sensing systems, such as strain gauge FlexForce [292] or the liquid metal-embedded elastomer sensor [293] can be used in glove-like devices; but even they can be too rigid to conform to the contours of the hand, resulting in limitations on natural motion during attempts at fine manipulative action. Recently, Liu *et al.* [294] introduced Velostat, a soft piezoresistive conductive film whose resistance changes under pressure. They used this material in an inertial measurement unit (IMU)-based position-sensing glove to reliably record manipulation demonstrations with fine-grained force information. This kind of measurement is particularly important for teaching systems to perform tasks with visually latent changes.

Consider the task of opening a medicine bottle with a child-safety locking mechanism. These bottles require the user to push or squeeze in specific places to unlock the cap. By design, attempts to open these bottles using a standard procedure will result in failure. Even if an agent visually observes a successful demonstration, attempted direct imitation will likely omit critical steps in the procedure, as the visual appearance of opening both medicine and traditional bottles are typically very similar if not identical. By using the Velostat [294] glove in demonstration, the fine forces used to unlock the child-safety mecha-

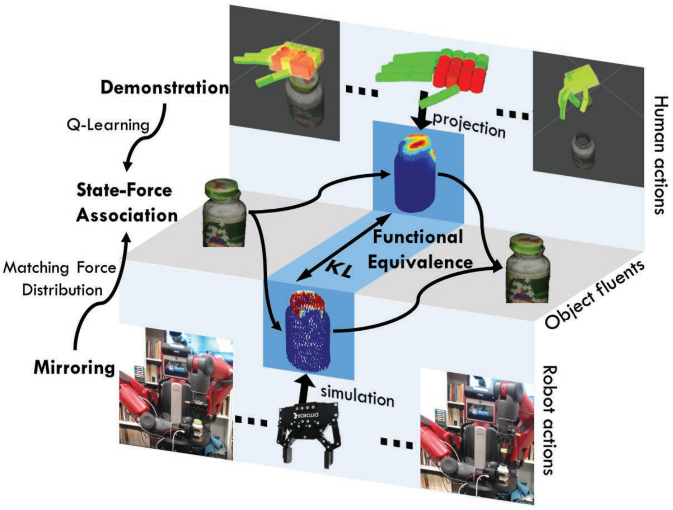


Figure 27: A robot mirrors human demonstrations with functional equivalence by inferring the action that produces similar force, resulting in similar changes in physical states. Q-learning is applied to similar types of forces with categories of object state changes to produce human–object-interaction (*hoi*) units. KL: Kullback–Leibler divergence. Reproduced from Ref. [298] with permission of Association for the Advancement of Artificial Intelligence, © 2019.

nism become observable. From these observations, Edmonds *et al*. [295, 296] taught 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 using the observed poses and forces. These two inputs were combined during planning to select the next optimal action. An augmented reality (AR) interface was also developed on top of this work to improve system interpretability and allow for easy patching of robot knowledge [297].

One major limitation of the above work is that the robot's actions are predefined, and the underlying structure of the task is not modeled. Recently, Liu *et al*. [298] proposed a *mirroring* approach and a concept of *functional manipulation* that extends the current LfD through a physics-based simulation to address the correspondence problem; see Fig. 27 [298] for more details. Rather than over-imitating the motion trajectories of the demonstration, the robot is encouraged 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. This approach has three characteristics distinguishing it from the standard LfD. First, it is *force-based*: these tactile perception-enabled demonstrations capture a deeper understanding of the physical world that a robot interacts with beyond visually observable space, providing an extra dimension that helps address the correspondence problem. Second, it is *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 comprises the state changes caused by forces, independent of embodiments. Finally, this method uses *mirroring without overimitation*: in contrast to the classic LfD, a robot does not necessarily mimic every action in a human demonstration; instead, the robot reasons about the motion to achieve the goal states based on the learned grammar and simulated forces.

6. Perceiving Intent: The Sense of Agency

In addition to inanimate physical objects, we live in a world with a plethora of animate and goal-directed agents, whose agency implies the ability to perceive, plan, make decisions, and achieve goals. Crucially, such a sense of agency further entails (i) the *intentionality* [299] to represent a future goal state and equifinal variability [300] to be able to achieve the intended goal state with different actions across contexts; and (ii) the*rationality of actions* in relation to goals [301] to devise the most efficient possible action plan. The perception and comprehension of intent enable humans to better understand and predict the behavior of other agents and engage with others in cooperative activities with shared goals. The construct of intent, as a basic organizing principle guiding how we interpret one another, has been increasingly granted a central position within accounts of human cognitive functioning, and thus should be an essential component of future AI.

In Section 6.1, we start with a brief introduction to what constitutes the concepts of "agency," which are deeply rooted in humans as young as six months old. Next, in Section 6.2, we explain the *rationality* principle as the mechanism with which both infants and adults perceive animate objects as intentional beings. We then describe how intent prediction is related to action prediction in modern computer vision and machine learning, but is in fact much more than predicting action labels; see Section 6.3 for a philosophical perspective. In Section 6.4, we conclude this section by providing a brief review of the building blocks for intent in computer vision and AI.

6.1. The Sense of Agency

In the literature, theory of mind (ToM) refers to the ability to attribute mental states, including beliefs, desires, and intentions, to oneself and others [302]. Perceiving and understanding an agent's intent based on their *belief* and *desire* is the ultimate goal, since people largely act to fulfill intentions arising from their beliefs and desires [303].

Evidence from developmental psychology shows that six-month-old infants see human activities as goal-directed behavior [304]. 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 or muscle movements [305, 306]. After their first birthday, infants begin to understand that an actor may consider various plans to pursue a goal, and choose one to intentionally enact based on environmental reality [307]. Eighteen-month-old children are able to both *infer* and *imitate* the intended goal of an action even if the action repeatedly fails to achieve the goal [308]. Moreover, infants can imitate actions in a rational, efficient way based on an evaluation of the action's situational constraints instead of merely copying movements, indicating that infants have a deep understanding of relationships among the environment, action, and underlying intent [309]. Infants can also perceive intentional relationships at varying levels of analysis, including concrete action goals, higher order plans, and collaborative goals [310].

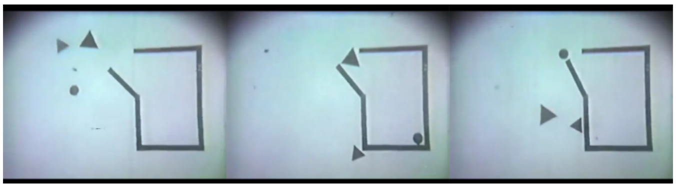


Figure 28: The seminal Heider–Simmel experiment [313]. Adults can perceive and attribute mental states from nothing but the motion of simple geometric shapes.

Despite the complexity of the behavioral streams we actually witness, we readily process action in intentional terms from infancy onward [303]. It is underlying *intent*, rather than surface behavior, that matters when we observe motions. One latent intention can make several highly dissimilar movement patterns conceptually cohesive. Even an identical physical movement could have a variety of different meanings depending on the intent motivating it; for example, the underlying intent driving a reach for a cup could be to either fill the cup or clean it. Thus, inference about others' intentions is what gives an observer the "gist" of human actions. Research has found that we do not encode the complete details of human motion in space; instead, we perceive motions in terms of intent. It is the constructed understanding of actions in terms of the actors' goals and intentions that humans encode in memory and later retrieve [303]. Reading intentions has even led to species-unique forms of cultural learning and cognition [307]. From infants to complex social institutions, our world is constituted of the intentions of its agents [311, 312, 307].

6.2. *From Animacy to Rationality*

Human vision has the uniquely social function of extracting latent mental states about goals, beliefs, and intentions from nothing but visual stimuli. Surprisingly, such visual stimuli do not need to contain rich semantics or visual features. An iconic illustration of this is the seminal Heider-Simmel display created in the 1940s [313]; see Fig. 28 for more detail. Upon viewing the 2D motion of three simple geometric shapes roaming around a space, human participants acting without any additional hints automatically and even irresistibly perceive "social agents," with a set of rich mental states such as goals, emotions, personalities, and coalitions. These mental states come together to form a story-like description of what is happening in the display, such as a hero saving a victim from a bully. Note that in this experiment, where no specific directions regarding perception of the objects were provided, participants still tended to describe the objects as having different sexes and dispositions. Another crucial observation is that human participants always reported the animated objects as "opening" or "closing" the door, similar to in Michotte's "entrance" display [79]; the movement of the animated object is imparted to the door through prolonged contact rather than through sudden impact. This interpretation of simple shapes as animated beings was a remarkable demonstration of how human vision is able to extract rich social relationships and mental states from sparse, symbolized inputs with extremely minimal visual features.

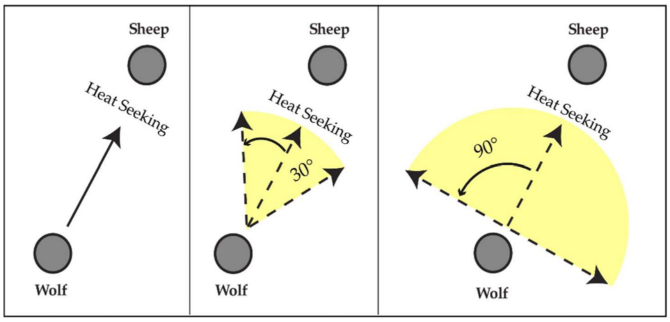


Figure 29: An illustration of *chasing subtlety* manipulation in the "Don't Get Caught" experiment. When chasing subtlety is set to zero, the wolf always heads directly toward the (moving) sheep in a "heat-seeking" manner. When the chasing subtlety is set to 30, the wolf always moves in the general direction of the sheep, but is not on a perfect, heat-seeking trajectory; instead, it can move in any direction within a 60-degree window that is always centered on the moving sheep. When the chasing subtlety is set to 90, the wolf's movement is even less directed; now the wolf may head in an orthogonal direction to the (moving) sheep, though it can still never move away from it. Reproduced from Ref. [317] with permission of Elsevier Inc., © 2009

In the original Heider-Simmel display, it is unclear whether the demonstrated visual perception of social relationships and mental states was attributable more or less to the dynamic motion of the stimuli, or to the relative attributes (size, shape, *etc.*) of the protagonists. Berry and Misovich [314] designed a quantitative evaluation of these two confounding variables by degrading the structural display while preserving its original dynamics. They reported a similar number of anthropomorphic terms as in the original design, indicating that the display's structural features are not the critical factors informing human social perception; this finding further strengthened the original finding that human perception of social relationships goes beyond visual features. Critically, when Berry and Misovich used static frames in both the original and degraded displays, the number of anthropomorphic terms dropped significantly, implying that the dynamic motion and temporal contingency were the crucial factors for the successful perception of social relationships and mental states. This phenomenon was later further studied by Bassili [315] in a series of experiments.

Similar simulations of biologically meaningful motion sequences were produced by Dittrich and Lea [316] in simple displays of moving letters. Participants were asked to identify one letter acting as a "wolf" chasing another "sheep" letter, or a "lamb" letter trying to catch up with its mother. These scholars' findings echoed the Heider-Simmel experiment; motion dynamics played an important factor in the perception of intentional action. Specifically, intentionality appeared stronger when the "wolf/lamb" path was closer to its target, and was more salient when the speed difference between the two was significant. Furthermore, Dittrich and Lea failed to find significantly different effects when the task was described in neutral terms (letters) in comparison with when it was described in intentional terms (*i.e.*, wolf/sheep).

Taken together, these experiments demonstrate that even the simplest moving shapes are irresistibly perceived in an intentional and goal-directed "social" way—through a holistic understanding of the events as an unfolding story whose charac-

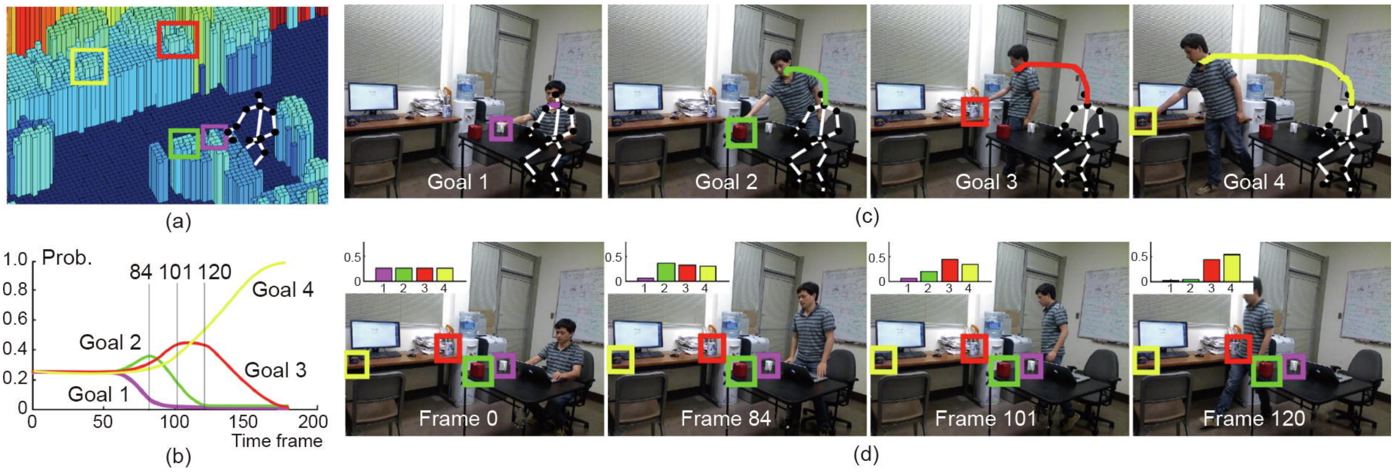


Figure 30: The plan inference task presented in Ref. [318], seen from the perspective of an observing robot. (a) Four different goals (target objects) in a 3D scene. (b) One outcome of the proposed method: the marginal probability (Prob.) of each terminal action over time. Note that terminal actions are marginal probabilities over the probability density described by the hierarchical graphical model. (c) 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; Goals 3 and 4 require standing up, moving, and reaching for different objects. (d) A progression of time corresponding to the results shown in (b). 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. Reproduced from Ref. [318] with permission of IEEE, © 2016.

ters have goals, beliefs, and intentions. A question naturally arises: what is the underlying mechanism with which the human visual system perceives and interprets such a richly social world? One possible mechanism governing this process that has been proposed by several philosophers and psychologists is the intuitive agency theory, which embodies the so-called "rationality principle." This theory states that humans view themselves and others as *causal* agents: (i) they devote their *limited* time and resources only to those actions that change the world in accordance with their intentions and desires; and (ii) they achieve their intentions *rationally* by maximizing their *utility* while minimizing their *costs*, given their *beliefs* about the world [319, 301, 320].

Guided by this principle, Gao *et al.* [317] 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 was to detect a hidden "wolf" and keep away from it for 20 s. The effectiveness of the wolf's chasing was measured by the percentage of the human's escape attempts that failed. Across trials, the wolf's pursuit strategy was manipulated by a variable called *chasing subtlety*, which controlled the maximum deviation from the perfect heat-seeking trajectory; see Fig. 29 [317] for more details. The results showed that humans can effectively detect and avoid wolves with small subtlety values, whereas wolves with modest subtlety values turned out to be the most "dangerous." A dangerous wolf can still approach a sheep relatively quickly; meanwhile, deviation from the most efficient heat-seeking trajectory severely disrupts a human's perception of being chased, leaving the crafty wolf undetected. In other words, they can effectively stalk the human-controlled "sheep" without being noticed. This result is consistent with the "rationality principle," where human perception assumes that an agent's intentional action will be one that maximizes its efficiency in reaching its goal.

Not only are adults sensitive to the cost of actions, as demonstrated above, but 6-to-12-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 than when a shorter route is available [321, 322]. Crucially, infants interpret actions as directed toward goal objects, looking longer when an agent reaches for a new object, even if the reach follows a familiar path [304]. Recently, Liu *et al.* [320] performed five looking-time experiments in which three-month-old 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 than when they are 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 [104, 323, 114]. Planning is a process by which intent causes action. Inverse planning, by inverting the rational planning model via Bayesian inference that integrates the likelihood of observed actions with prior mental states, can infer the latent mental intent. Based on inverse planning, Baker *et al.* [104] proposed a framework for goal inference, in which the bottom-up information of behavior observations and the top-down prior knowledge of goal space are integrated to allow inference of underlying intent. In addition, Bayesian networks, with their flexibility in representing probabilistic dependencies and causal relationships, as well as the efficiency of inference methods, have proven to be one of the most powerful and successful approaches for intent recognition [324, 325, 326, 323].

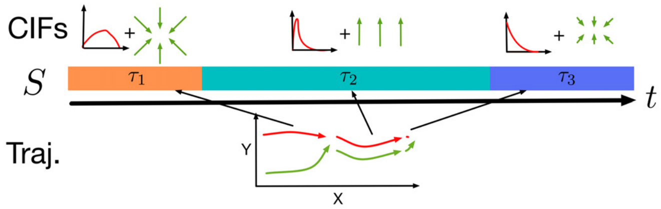


Figure 31: Inference of human interaction from motion trajectories. The top row demonstrates change within a conditional interactive field (CIF) in sub-interactions as the interaction proceeds, where the CIF models the expected relative motion pattern conditioned on the reference agent's motion. The bottom illustrates the change in interactive behaviors in terms of motion trajectories (Traj.). The colored bars in the middle depict the types of sub-interactions (S). Reproduced from Ref. [112] with permission of Cognitive Science Society, Inc., © 2017.

Moving from the symbolic input to real video input, Holtzen *et al*. [318] presented an inverse planning method to infer human hierarchical intentions from partially observed RGB-D videos. Their algorithm is able to infer human intentions by reverse-engineering decision-making and action planning processes in human minds under a Bayesian probabilistic programming framework; see Fig. 30 [318] for more details. The intentions are represented as a novel hierarchical, compositional, and probabilistic graph structure that describes the relationships between actions and plans.

By bridging from the abstract Heider-Simmel display to aerial videos, Shu *et al*. [112] proposed a method to infer humans' intentions with respect to interaction by observing motion trajectories (Fig. 31). A non-parametric exponential potential function is taught to derive "social force and fields" through the calculus of variations (as in Landau physics); such force and fields explain human motion and interaction in the collected drone videos. The model's results fit well with human judgments of propensity or inclination to interact, and demonstrate the ability to synthesize decontextualized animations that have a controlled level of interactiveness.

In outdoor scenarios, Xie *et al*. [72] jointly inferred object functionality and human intent by reasoning about human activities. Based on the rationality principle, the people in the observed videos are expected to intentionally take the shortest possible paths toward functional objects, subject to obstacles, that allow the people to satisfy certain of their needs (*e.g*., a vending machine can quench thirst); see Fig. 9. Here, the functional objects are "dark matter" since they are typically difficult to detect in low-resolution surveillance videos and have the functionality to "attract" people. Xie *et al*. formulated agent-based Lagrangian mechanics wherein human trajectories are probabilistically modeled as motions in many layers of "dark energy" fields, and wherein each agent can choose to allow a particular force field to affect its motions, thus defining the minimum-energy Dijkstra path toward the corresponding "dark matter" source. 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' intentions.

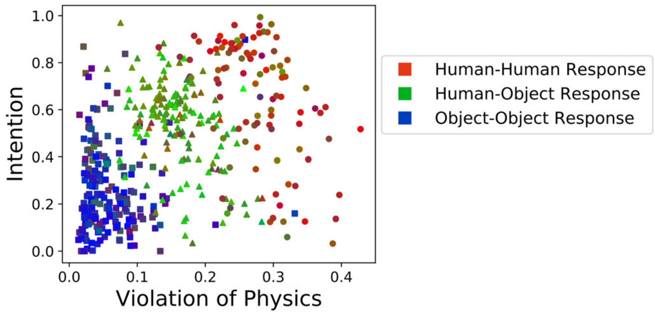


Figure 32: Constructed psychological space including human-human (HH) animations with 100% animacy degree, human–object (HO) animations, and object-object (OO) animations. Here, a stimulus is depicted by a data point with coordinates derived by the model, and the colors of the data points indicate the average human responses to this stimulus. The two variables in the space are the average of the measures of the degree of violation of physical laws and the values indicating the presence of intent between two entities. The shapes of data points correspond to the interaction types used in the simulation for generating the corresponding stimuli (circle: HH, triangle: HO, square: OO). Reproduced from Ref. [113] with permission of Cognitive Science Society, Inc., © 2019.

6.3. *Beyond Action Prediction*

In modern computer vision and AI systems [327], intent is related to action prediction *much* more profoundly than through simply predicting action labels. Humans have a strong and early-emerging inclination to interpret actions in terms of intention as part of a long-term process of *social learning* about novel means and novel goals. From a philosophical perspective, Csibra *et al*. [103] contrasted three distinct mechanisms: (i) action-effect association, (ii) simulation procedures, and (iii) 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 intent representations [328]; and that such simulation processes are at the heart of the development of intentional action interpretation [308]. In order to understand others' intentions, humans subconsciously empathize with the person they are observing and estimate what their own actions and intentions might be in that situation. Here, action-effect association [329] plays an important role in quick online intent prediction, and the ability to encode and remember these two component associations contributes to infants' imitation skills and intentional action understanding [330]. Accumulating neurophysiological evidence supports such simulations in the human brain; one example is the mirror neuron [331], which has been linked to intent understanding in many studies [332, 102]. 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*., Ref. [333]), which poses challenges to the simulation theory of intent attribution.

To address social learning, a teleological action interpretational system [335] takes a "functional stance" for the com-

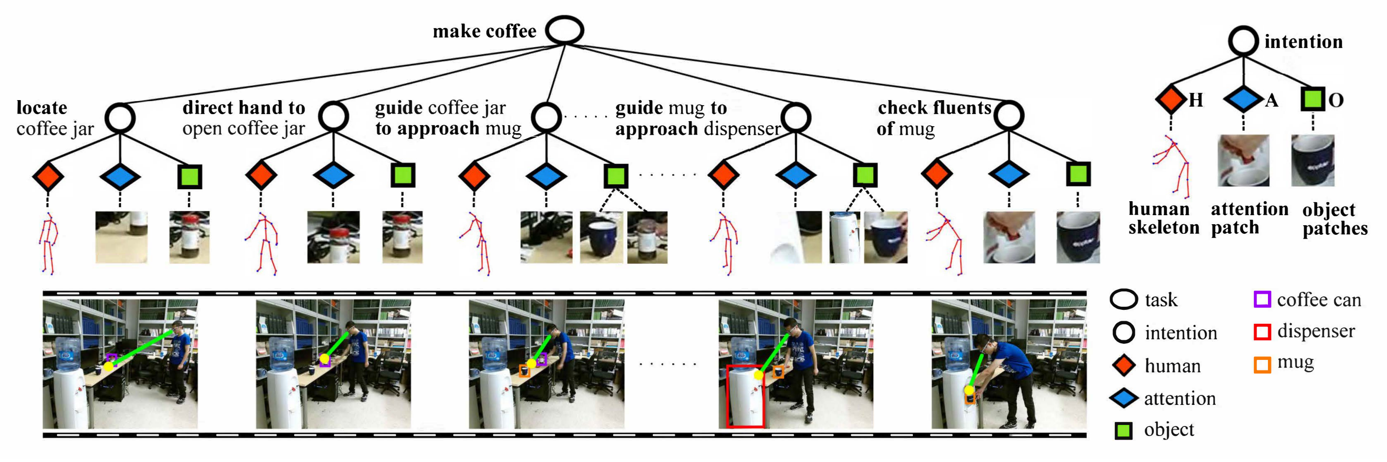


Figure 33: A task is modeled as sequential intentions in terms of hand-eye coordination with a human-attention-object (HAO) graph. Here, an intention is represented through inverse planning, in which human pose, human attention, and a visible object provide context with which to infer an agent's intention. Reproduced from Ref. [334] with permission of the authors, © 2018.

putational representation of goal-directed action [103], where such teleological representations are generated by the aforementioned inferential "rationality principle" [336]. In fact, the very notion of "action" implies motor behavior performed by an agent that is conceived in relation to the end state that agent wants to achieve. Attributing a goal to an observed action enables humans to predict the course of future actions, evaluate causal efficacy or certain actions, and justify an action itself. Furthermore, action predictions can be made by breaking down a path toward a goal into a hierarchy of sub-goals, the most basic of which are comprised of elementary motor acts such as grasping.

These three mechanisms do not compete; instead, they 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 Intent in Computer Vision

Understanding and predicting human intentions from images and videos is a research topic that is driven by many real-world applications, including visual surveillance, human-robot interaction, and autonomous driving. In order to better predict intent 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 relationships, and communicative gestures (*e.g.*, pointing).

Motion trajectory alone could be a strong signal for intent prediction, as discussed in Section 6.2. With intuitive physics and perceived intent, humans also demonstrate the ability to distinguish social events from physical events with very limited motion trajectory stimuli, such as the movements of a few simple geometric shapes. Shu *et al.* [113] studied possible 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 important dimensions: (i) an intuitive sense of whether physical laws are obeyed or violated, and (ii) an impression of whether an agent possesses intent as inferred from the movements of simple shapes; see Fig. 32 [113]. Their experiments demonstrate that the constructed psychological space successfully partitions human perception of physical versus social events.

Eye gaze, being closely related to underlying attention, intent, emotion, personality, and anything a human is thinking and doing, also plays an important role in allowing humans to "read" other peoples' minds [337]. Evidence from psychology suggests that eyes are a cognitively special stimulus with distinctive, "hardwired" pathways in the brain dedicated to their interpretation, revealing humans' unique ability to infer others' intent from eye gazes [338]. Social eye gaze functions also transcend cultural differences, forming a kind of universal language [339]. Computer vision and AI systems heavily rely on gazes as cues for intent prediction based on images and videos. For example, the system developed by Wei *et al.* [334] jointly inferred human attention, intent, and tasks from videos. Given an RGB-D video in which a human performs a task, the system answered three questions simultaneously: (i) "Wwere is the human looking?"—attention/gaze prediction; (ii) "why is the human looking?"—intent prediction; and (iii) "what task is the human performing?"—task recognition. Wei *et al.* [334] proposed a hierarchical human-attention-object (HAO) model that represents tasks, intentions, and attention under a unified framework. Under this model, a task is represented as sequential intentions described by hand-eye coordination under a planner represented by a grammar; see Fig. 33 for details [334].

Communicative gazes and gestures (*e.g.*, pointing) stand out for intent expression and perception in collaborative interactions. Humans need to recognize their partners' communicative intentions in order to collaborate with others and success-

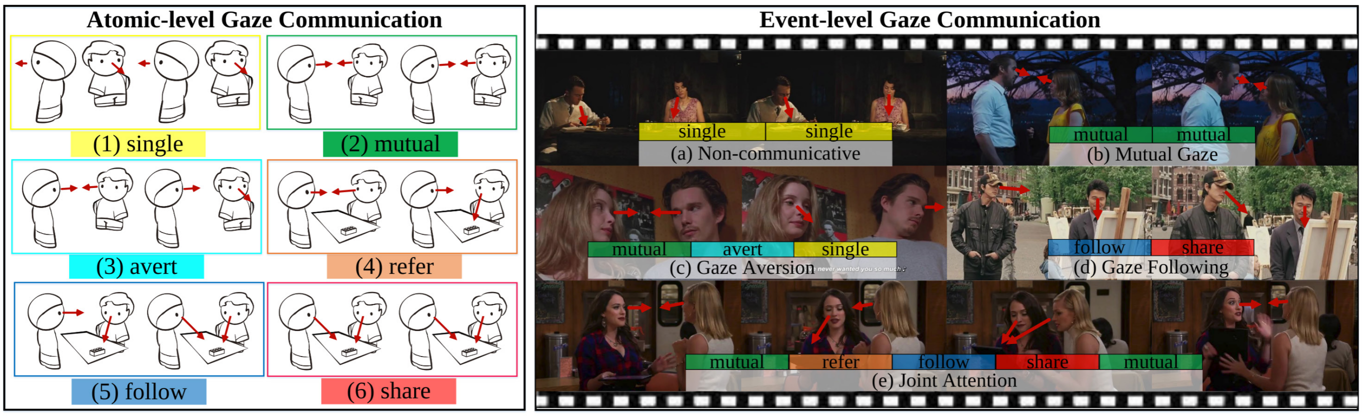


Figure 34: Human gaze communication dynamics on two hierarchical levels: (i) Atomic-level gaze communication describes the fine-grained structures in human gaze interactions; and (ii) event-level gaze communication refers to long-term social communication events temporally composed of atomic-level gaze communications. Reproduced from Ref. [340] with permission of the authors, © 2019.

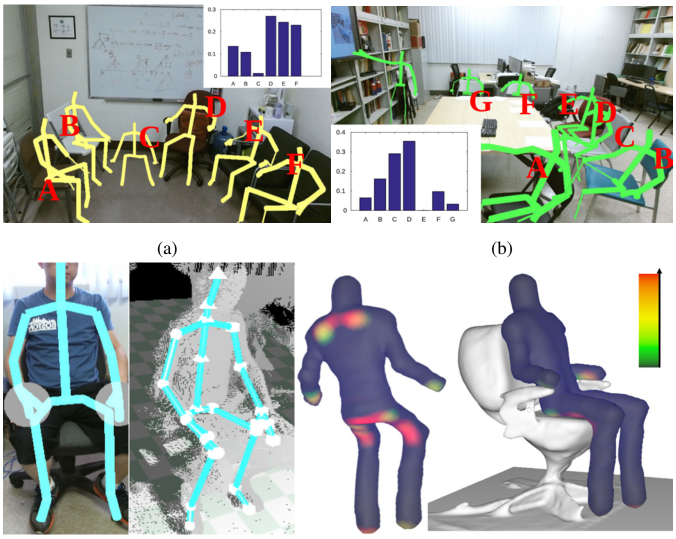
fully survive in the world. Human communication in mutualistic collaboration often involves agents informing recipients of things they believe will be useful or relevant to them. Melis and Tomasello *et al.* [341] investigated whether pairs of chimpanzees were capable of communicating to ensure coordination during collaborative problem-solving. In their experiments, the chimpanzee pairs needed two tools to extract fruit from an apparatus. The communicator in each pair could see the location of the tools (hidden in one of two boxes), but 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 fruit. As demonstrated by this study, even chimpanzees have obtained the necessary socio-cognitive skills to naturally develop a simple communicative strategy to ensure coordination in a collaborative task. To model such a capability that is demonstrated in both chimpanzees and humans, Fan *et al.* [342] studied the problem of human communicative gaze dynamics. They examined the inferring of shared eye gazes in third-person social scene videos, which is a phenomenon in which two or more individuals simultaneously look at a common target in social scenes. A follow-up work [340] studied various types of gaze communications in social activities from both the atomic level and event level (Fig. 34). A spatiotemporal graph network was proposed to explicitly represent the diverse interactions in the social scenes and to infer atomic-level gaze communications.

Humans communicate intentions multimodally; thus, facial expression, head pose, body posture and orientation, arm motion, gesture, proxemics, and relationships with other agents and objects can all contribute to human intent analysis and comprehension. Researchers in robotics try to equip robots with the ability to act "naturally," or to be subject to "social affordance," which represents action possibilities that follow basic social norms. Trick *et al.* [343] proposed an approach for multimodal intent recognition that focuses on uncertainty reduction through classifier fusion, considering four modalities: speech, gestures, gaze directions, and scene objects. Shu *et al.* [344] presented 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 by learning structural representations of human-human (HH) and human-object-human (HOH) interactions that describe how agents' body parts move and what spatial relationships they should maintain in order to complete each sub-goal, a robot can infer what its own movement should be in reaction to the motion of the human body. Such social affordance could also be represented by a hierarchical grammar model [345], enabling real-time motion inference for human-robot interaction; the learned model was demonstrated to successfully infer human intent and generate humanlike, socially appropriate response behaviors in robots.

7. Learning Utility: The Preference of Choices

Rooted in 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 their beliefs and desires to maximize its expected utility. This is known as the principle of maximum expected utility. We argue that the majority of the observational signals we encounter in daily life are driven by this simple yet powerful principle—an invisible "dark" force that governs the mechanism that explicitly or implicitly underlies human behaviors. Thus, studying utility could provide a computer vision or AI system with a deeper understanding of its visual observations, thereby achieving better generalization.

According to the classic definition of utility, the utility that a decision-maker gains from making a choice is measured with a utility function. A utility function is a mathematical formulation that ranks the preferences of an individual such that $U(a) > U(b)$, where choice a is preferred over 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 func-



(a) (b)

(c) (d)

Figure 35: Examples of sitting in (a) an office and (b) a meeting room. In addition to geometry and appearance, people consider other important factors when deciding where to sit, including comfort level, reaching cost, and social goals. The histograms indicate human preferences for different candidate chairs. Based on these observations, it is possible to infer human utility during sitting from videos[233]. (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. Reproduced from Ref. [233] with permission of the authors, © 2016.

tion in its own deliberations. By observing a rational agent's preferences, however, an observer can construct a utility function that represents what the agent is actually trying to achieve, even if the agent does not know it [346]. It is also worth noting that 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 how people *should* behave; such a distinction is crucial for the discipline of economics, and for the devising of algorithms and systems to interpret observational signals.

Although Jeremy Bentham [117] is often regarded as the first scholar to systematically study utilitarianism—the philosophical concept that was later borrowed by economics and game theory, the core insight motivating the theory was established much earlier by Francis Hutcheson [347] on action choice. In the field of philosophy, utilitarianism is considered a normative ethical theory that places the locus of right and wrong solely on the outcomes (consequences) of choosing one action/policy over others. As such, it moves beyond the scope of one's own interests and takes into account the interests of others [347, 348]. The term has been adopted by the field of economics, where a utility function represents a consumer's order of preferences given a set of choices. As such, the term "utility" is now devoid of its original meaning.

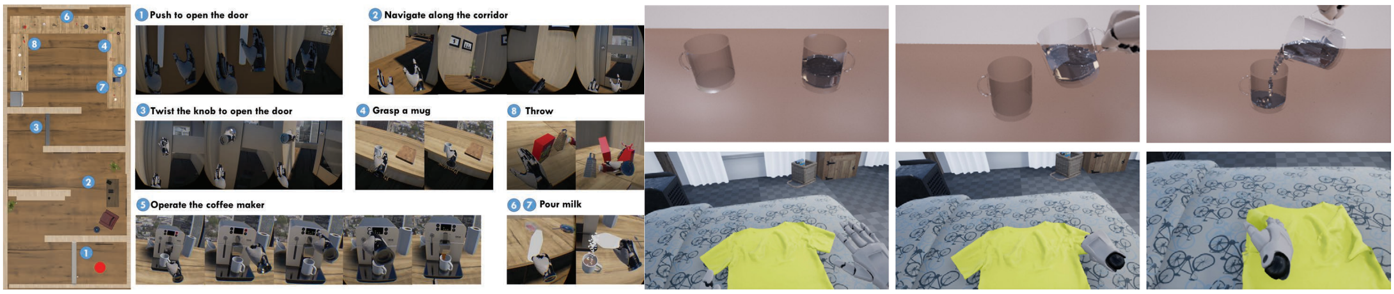
Formally, the core idea behind utility theory is straightforward: every possible action or state within a given model can be described with a single, uniform value. This value, usually referred to as *utility*, describes the usefulness of that action within the given context. Note that the concept of *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 four recent case studies in computer vision, robotics, linguistics, and social learning that use a utility-driven learning approach.

As shown in Fig. 35 [233], by observing the choices people make in videos (particularly in selecting a chair on which to sit), a computer vision system [233] is able to learn the comfort intervals of the forces exerted on different body parts while sitting, thereby accounting for people's preferences in terms of human *internal* utility.

Similarly, Shukla *et al*. [350] adopted the idea of learning human utility in order to teach a robotics task using human demonstrations. A proof-of-concept work shows a pipeline in which the agent learns the *external* utility of humans and plans a cloth-folding task using this learned utility function. Specifically, under the assumption that the utility of the goal states is higher than that of the initial states, this system learns the *external* utility of humans by ranking pairs of states extracted from images.

In addition, the rationality principle has been studied in the field of linguistics and philosophy, notably in influential work on the theory of implicature by Grice [351]. The core insight of Grice's work is that language use is a form of rational action; thus, technical tools for reasoning about rational action should elucidate linguistic phenomena [352]. Such a goal-directed view of language production has led to a few interesting language games [353, 354, 355, 356, 357, 358], the development of engineering systems for natural language generation [359], and a vocabulary for formal descriptions of pragmatic phenomena in the field of game theory [360, 361]. More recently, by assuming the communications between agents to be helpful yet parsimonious, the "Rational Speech Act" [362, 352] model has demonstrated promising results in solving some challenging referential games.

By materializing the internal abstract social concepts using external explicit forms, utility theory also plays a crucial role in social learning, and quantizes an actor's belief distribution. Utility, which is analogous to the "dark" currency circulating in society, aligns social values better among and within groups. By modeling how people value the decision-making process as permissible or not using utilities, Kleiman-Weiner *et al*. [363] were able to solve challenging situations with social dilemma. Based on how the expected utility influences the distribution, social goals (*e.g.*, cooperation and competition) [364, 365] and fairness [366] can also be well explained. On a broader scale, utility can enable individuals to be self-identified in society during the social learning process; for example, when forming basic social concepts and behavior norms during the early stages of the development, children compare their own meta-values with the observed values of others [367].


1 Push to open the door
2 Navigate along the corridor
3 Twist the knob to open the door
4 Grasp a mug
8 Throw
5 Operate the coffee maker
6 7 Pour milk

(a) Various task executions in VRGym

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

Figure 36: VRGym, an example of a 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 and evaluate the success of task execution; (b) in addition to the rigid-body simulation, VRGym supports realistic real-time fluid and cloth simulations, leveraging state-of-the-art game engines. Reproduced from Ref. [349] with permission of Association for Computing Machinery, © 2019.

8. Summary and Discussions

Robots are mechanically capable of performing a wide range of complex activities; however, in practice, they do very little that is useful for humans. Today's robots fundamentally lack physical and social common sense; this limitation inhibits their capacity to aid in our daily lives. In this article, we have reviewed five concepts that are the crucial building blocks of common sense: functionality, physics, intent, causality, and utility (FPICU). We argued that these cognitive abilities have shown potential to be, in turn, the building blocks of cognitive AI, and should therefore be the foundation of future efforts in constructing this cognitive architecture. The positions taken in this article are not intended to serve as *the* solution for the future of cognitive AI. Rather, by identifying these crucial concepts, we want to call attention to pathways that have been less well explored in our rapidly developing AI community. There are indeed many other topics that we believe are also essential AI ingredients; for example:

- *A physically realistic VR/MR platform: from big data to big tasks.* Since FPICU is "dark"—meaning that it often does not appear in the form of pixels—it is difficult to evaluate FPICU in traditional terms. Here, we argue that the ultimate standard for validating the effectiveness of FPICU in AI is to examine whether an agent is capable of (i) accomplishing the very same task using different sets of objects with different instructions and/ or sequences of actions in different environments; and (ii) rapidly adapting such learned knowledge to entirely new tasks. By leveraging state-of-the-art game engines and physics-based simulations, we are beginning to explore this possibility on a large scale; see Section 8.1.
- *Social system: the emergence of language, communication, and morality.* While FPICU captures the core components of a single agent, modeling interaction among and within agents, either in collaborative or competitive situations [368], is still a challenging problem. In most cases, algorithms designed for a single agent would be difficult to generalize to a multiple-agent systems (MAS) setting [369, 370, 371]. We provide a brief review of three related topics in Section 8.2.
- *Measuring the limits of an intelligence system: IQ tests.* Studying FPICU opens a new direction of analogy and relational reasoning [372]. Apart from the four-term analogy (or proportional analogy), John C. Raven [373] proposed the raven's prograssive matrices test (RPM) in the image domain. The RAVEN dataset [374] was recently introduced in the computer vision community, and serves as a systematic benchmark for many 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 in 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" [375]. To reach this goal, we have seen the increasing use of synthetic data and simulation platforms for indoor scenes in recent years by leveraging state-of-the-art game engines and free, publicly available 3D content [376, 377, 288, 378], including MINOR [379], HoME [380], Gibson [381], House3D [382], AI-THOR [383], VirtualHome [384], VRGym [349] (Fig. 36), and VRKitchen [385]. In addition, the AirSim [386] open-source simulator was developed for outdoor scenarios. Such synthetic data could be relatively easily scaled up compared with traditional data collection and labeling processes. With increasing realism and faster rendering speeds built on dedicated hardware, synthetic data from the virtual world is becoming increasingly similar to data collected from the physical world. In these realistic virtual environments, it is possible to evaluate any AI method or system from a much more holistic perspective. Using a holistic evaluation, whether a method or a system is intelligent or not is no longer measured by the successful performance of a single narrow task; rather, it is measured by the ability to perform well across various tasks: the perception of environments, planning of actions, predictions of other agents' behaviors, and ability to rapidly adapt learned knowledge to new environments for new tasks.

To build this kind of task-driven evaluation, physics-based simulations for multi-material, multi-physics phenomena (Fig. 37) will play a central role. We argue that cognitive AI needs to accelerate the pace of its adoption of more advanced

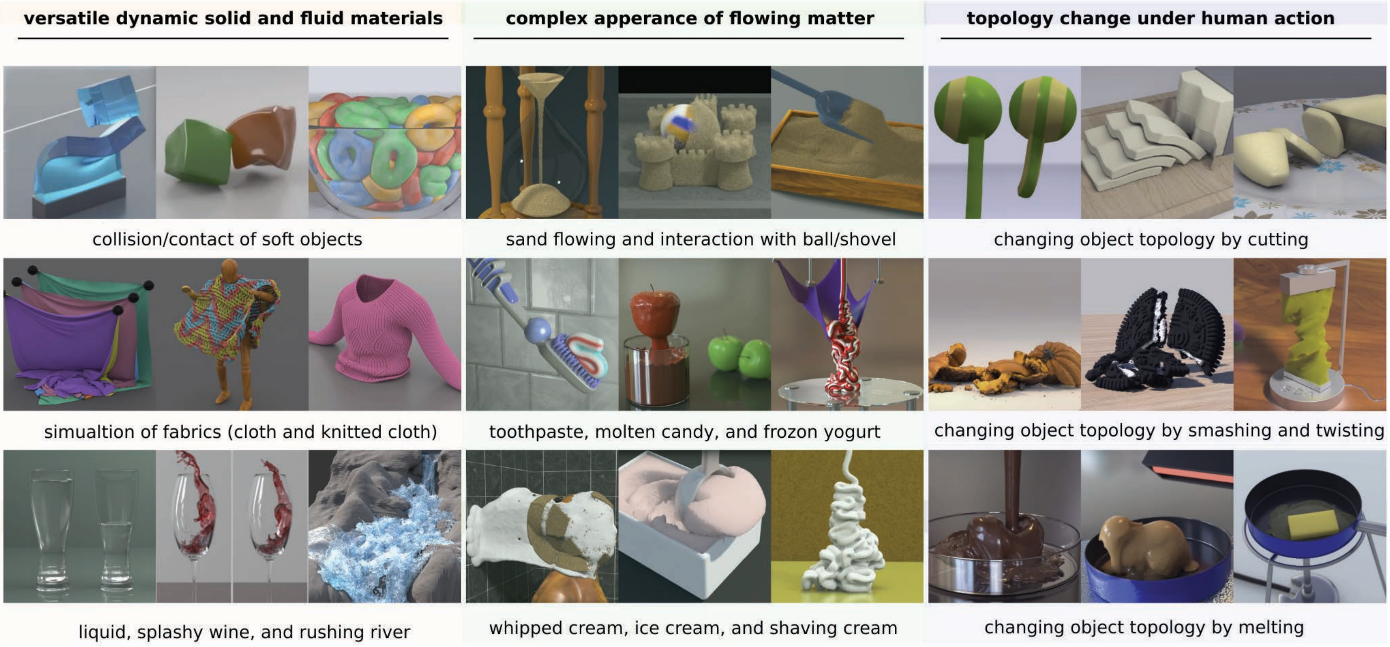

versatile dynamic solid and fluid materials
complex apperance of flowing matter
topology change under human action
collision/contact of soft objects
sand flowing and interaction with ball/shovel
changing object topology by cutting
simualtion of fabrics (cloth and knitted cloth)
toothpaste, molten candy, and frozon yogurt
changing object topology by smashing and twisting
liquid, splashy wine, and rushing river
whipped cream, ice cream, and shaving cream
changing object topology by melting

Figure 37: Diverse physical phenomena simulated using the material point method (MPM).

simulation models from computer graphics, in order to benefit from the capability of highly predictive forward simulations, especially graphics processing unit (GPU) optimizations that allow real-time performance [387]. Here, we provide a brief review of the recent physics-based simulation methods, with a particular focus on the material point method (MPM).

The accuracy of physics-based reasoning greatly relies on the fidelity of a physics-based simulation. Similarly, the scope of supported virtual materials and their physical and interactive properties directly determine the complexity of the AI tasks involving them. Since the pioneering work of Terzopoulos*et al*. [388, 389] for solids and that of Foster and Metaxas [390] for fluids, many mathematical and physical models in computer graphics have been developed and applied to the simulation of solids and fluids in a 3D virtual environment.

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

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 [391, 392]. Eulerian methods are more error-prone and require delicate treatment when dealing with deforming material interfaces and boundary conditions, since no explicit tracking of them is available.
2. Lagrangian mesh-based methods, represented by FEM [393, 394, 395], 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 acfem is mesh distortion and lack of contact during large deformations [396, 397] or topologically changing events [398].
3. Lagrangian mesh-free methods, such as smoothed particle hydrodynamics (SPH) [399] and the reproducing kernel particle method (RKPM) [400]. These methods allow arbitrary deformation but require expensive operations such as neighborhood searching [401]. Since the interpolation kernel is approximated with neighboring particles, these methods also tend to suffer from numerical instability issues.
4. Hybrid Lagrangian–Eulerian methods, such as the arbitrary Lagrangian–Eulerian (ALE) methods [402] and the MPM. These methods (particularly the MPM) combine the 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 [403, 404] from computational fluid dynamics to computational solid mechanics, the MPM has proven to be a promising discretization choice for simulating many solid and fluid materials since its introduction two decades ago [405, 406]. In the field of visual computing, existing work includes snow [407, 408], foam [409, 410, 411], sand [412, 413], rigid body [414], fracture [415, 416], cloth [417], hair [418], water [419], and solid-fluid mixtures [420, 421, 422]. In computational engineering science, this method 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 [423], material failure [424, 425], hyper-velocity impact [426, 427], explosion [428], fluid-structure interaction [429, 430], biomechanics [431], geomechanics [432], and many other examples that are considerably more difficult when addressed with traditional, non-hybrid approaches. In addition to experiencing a tremendously expanding scope of application, the MPM's discretization scheme has been extensively improved [433]. To alleviate numerical inaccuracy and stability issues associated with the original MPM formulation, researchers have proposed different variations of the MPM, including the generalized interpolation material point (GIMP) method [434, 435], the convected particle domain interpolation (CPDI) method [436], and the dual domain material point (DDMP) method [437].

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 [368]). To scale up from rule-based systems, decentralized partially observable Markov decision processes were devised to model multi-agent interaction, with communication being considered as a special type of action [438, 439]. As with the success of RL in single-agent games [440], generalizing Q-learning [441, 371] and actor-critic [369, 442]-based methods from single-agent system to MAS have 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 [370, 443, 444] has shown that agents can learn how to communicate with continuous signals that are only decipherable within a group. The emergence of more realistic communication protocols using discrete messages has been explored in various types of communication games [445, 446, 447, 448], in which agents need to process visual signals and attach discrete tokens to attributes or semantics of images in order to form effective protocols. By letting groups of agents play communication games spontaneously, several linguistic phenomena in emergent communication and language have been studied [449, 450, 451].

Morality is an abstract and complex concept composed of common principles such as fairness, obligation, and permissibility. It is deeply rooted in the tradeoffs people make every day when these moral principles come into conflict with one another [452, 453]. 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: one might view it as corrupt, the other as a moral obligation [367]. Indeed, the same principle might be viewed differently in two social groups with distinct cultures [454]. Even within the same social group, different individuals might have different standards on the same moral principle or event that triggers moral judgment [455, 456, 457]. Many works have proposed theoretical accounts for categorizing the different measures of welfare used in moral calculus, including "base goods" and "primary goods" [458, 459], "moral foundations" [460], and the feasibility of value judgment from an infant's point of view [461]. Despite its complexity and diversity, devising a computational account of morality and moral judgment is an essential step on the path toward building humanlike machines. One recent approach to moral learning combines utility calculus and Bayesian inference to distinguish and evaluate different principles [367, 462, 363].

8.3. *Measuring the Limits of Intelligence System: IQ tests*

In the literature, we call two cases analogous if they share a common *relationship*. Such a relationship does not need to be among entities or ideas that use the same label across disciplines, such as computer vision and AI; rather, "analogous" emphasizes commonality on a more abstract level. For example, according to Ref. [463], the earliest major scientific discovery made through analogy can be dated back to imperial Rome, when investigators analogized waves in water and sound. They posited that sound waves and water waves share similar behavioral properties; for example, their intensities both diminish as they propagate across space. To make a successful analogy, the key is to understand *causes and their effects* [464].

The history of analogy can be categorized into three streams of research; see Ref. [372] for a capsule history and review of the literature. One stream is the psychometric tradition of four-term or "proportional" analogies, the earliest discussions of which can be traced back to Aristotle [465]. An example in AI is the *word2vec* model [466, 467], which is capable of making a four-term word analogy; for example, [king:queen::man:woman]. In the image domain, a similar test was invented by John C. Raven [373]—the raven's prograssive matrices test (RPM).

RPM has been widely accepted and is believed to be highly correlated with real intelligence [468]. Unlike visual question answering (VQA) [469], which lies at the periphery of the cognitive ability test circle [468], RPM lies directly at the center: it is diagnostic of abstract and structural reasoning ability [470], and captures the defining feature of high-level cognition—that is, *fluid intelligence* [471]. It has been shown that RPM is more difficult than existing visual reasoning tests in the following ways [374]:

- Unlike VQA, where natural language questions usually imply what the agent should pay attention to in an image, RPM relies merely on visual clues provided in the matrix. The *correspondence problem* itself, that is, the ability to find corresponding objects across frames to determine their relationship, is already a major factor distinguishing populations of different intelligence [468].

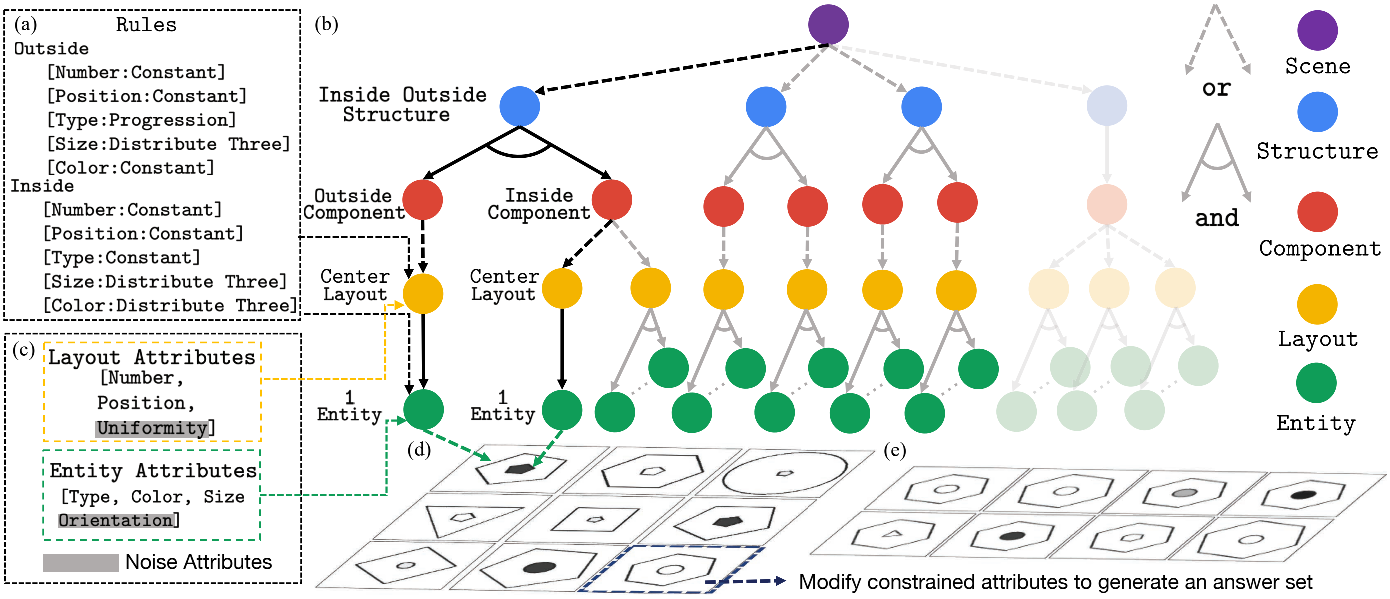


Figure 38: The RAVEN creation process proposed in Ref. [374]. A graphical illustration of (a) the grammar production rules used in (b) A-SIG. (c) Note that Layout and Entity have associated attributes. (d) A sample problem matrix and (e) a sample candidate set. Reproduced from Ref. [374] with permission of the authors, © 2019.

- 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 to understand *analogy*, and the grasp of *structure* must be taken into consideration in order to solve an 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 process extremely difficult.

The RAVEN dataset [374] was created to push the limit of current vision systems' reasoning and analogy-making ability, and to promote further research in this area. The dataset is designed to focus on reasoning and analogizing instead of only visual recognition. It is unique in the sense that it builds a semantic link between the visual reasoning and structural reasoning in RPM by grounding each problem into a sentence derived from an attributed stochastic image grammar attributed stochastic image grammar (A-SIG): each instance is a sentence sampled from a predefined A-SIG, and a rendering engine transforms the sentence into its corresponding image. (See Fig. 38 [374] for a graphical illustration of the generation process.) This semantic link between vision and structure representation opens new possibilities by breaking down the problem into image understanding and abstract-level structure reasoning. Zhang *et al.* [374] empirically demonstrated that models using a simple structural reasoning module to incorporate both vision-level understanding and abstract-level reasoning and analogizing notably improved their performance in RPM, whereas a variety of prior approaches to relational learning performed only slightly better than a random guess.

Analogy consists of more than mere spatiotemporal parsing and structural reasoning. For example, the *contrast effect* [472] has been proven to be one of the key ingredients in relational and analogical reasoning for both human and machine learning [473, 474, 475, 476, 477]. Originating from perceptual learning [478, 479], it is well established in the field of psychology and education [480, 481, 482, 483, 484] that teaching new concepts by comparing noisy examples is quite effective. Smith and Gentner [485] summarized that comparing cases facilitates transfer learning and problem-solving, as well as the ability to learn relational categories. In his structure-mapping theory, Gentner [486] postulated that learners generate a structural alignment between two representations when they compare two cases. A later article [487] firmly supported this idea and showed that finding the individual difference is easier for humans when similar items are compared. A more recent study from Schwartz *et al.* [488] also showed that contrasting cases helps to foster an appreciation of deep understanding. To retrieve this missing treatment of contrast in machine learning, computer vision and, more broadly, in AI, Zhang *et al.* [489] proposed methods of learning perceptual inference that explicitly introduce 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 at 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, which comprises ideas from contrast effects and perceptual inference, achieved state-of-the-art performance on major RPM datasets.

Parallel to work on RPM, work on *number sense* [490] bridges the induction of symbolic concepts and the competence of problem-solving; in fact, number sense could be regarded as a mathematical counterpart to the visual reasoning task of

RPM. A recent work approaches the analogy problem from this perspective of strong mathematical reasoning [491]. Zhang *et al.* [491] studied the machine number-sense problem and proposed a dataset of visual arithmetic problems for abstract and relational reasoning, where the machine is given two figures of numbers following hidden arithmetic computations and is tasked to work out a missing entry in the final answer. Solving machine number-sense problems is non-trivial: the system must both recognize a number and interpret the number with its contexts, shapes, and relationships (*e.g.*, symmetry), together with its proper operations. Experiments show that the current neural-network-based models do not acquire mathematical reasoning abilities after learning, whereas classic search-based algorithms equipped with an additional perception module achieve a sharp performance gain with fewer search steps. This work also sheds some light on how machine reasoning could be improved: the fusing of classic search-based algorithms with modern neural networks in order to discover essential number concepts in future research would be an encouraging development.

9. Acknowledgments

This article presents 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 (in alphabetical order: Carnegie Mellon University (CMU), Massachusetts Institute of Technology (MIT), Stanford University, University of California at Los Angeles (UCLA), University of Illinois at Urbana-Champaign (UIUC), and Yale University) and the UK (in alphabetical order: University of Birmingham, University of Glasgow, University of Leeds, and University of Oxford)[2]. The MURI team also holds an annual review meeting at various locations together with two related series of CVPR/CogSci workshops[3][4].

We are grateful to the editor of the special issue and the two reviewers for their valuable comments that have helped improve the presentation of the paper. We thank the following colleagues for helpful discussions on various sections: Professor Chenfanfu Jiang at the University of Pennsylvania; Dr. Behzad Kamgar-Parsi at the Office of Naval Research (ONR) and Dr. Bob Madahar at the Defence Science and Technology Laboratory (DSTL); Luyao Yuan, Shuwen Qiu, 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.

The work reported herein is supported by MURI ONR (N00014-16-1-2007), DARPA XAI (N66001-17-2-4029), and ONR (N00014-19-1-2153).

[2] See https://vcla.stat.ucla.edu/MURI_Visual_CommonSense/

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

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

References

[1] 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] 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.

[23] 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.

[24] 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.

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

[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] 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.

[38] Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, Song-Chun Zhu, Perspectivenet: 3d object detection from a single rgb image via perspective points, in: NeurIPS, 2019.

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

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

[41] 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.

[42] 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.

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

[44] 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.

[45] Elizabeth R Chrastil, William H Warren, From cognitive maps to cognitive graphs, PloS one 9 (11) (2014) e112544.

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

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

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

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

[50] Rudolf Karl Luneburg, Mathematical analysis of binocular vision, Princeton University Press, 1947.

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

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

[53] 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.

[54] 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.

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

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

[57] 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.

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

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

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

[61] 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 Representations (ICLR), 2019.

[62] Jianwen Xie, Ruiqi Gao, Erik Nijkamp, Song-Chun Zhu, Ying Nian Wu, Representation learning: A statistical perspective, Annual Review of Statistics and Its Application 7.

[63] 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.

[64] Jenny Vuong, Andrew Fitzgibbon, Andrew Glennerster, Human pointing errors suggest a flattened, task-dependent representation of space, bioRxiv (2018) 390088.

[65] Hoon Choi, Brian J Scholl, Perceiving causality after the fact: Postdiction in the temporal dynamics of causal perception, Perception 35 (3) (2006) 385–399.

[66] Brian J Scholl, Ken Nakayama, Illusory causal crescents: Misperceived spatial relations due to perceived causality, Perception 33 (4) (2004) 455–469.

[67] 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.

[68] Brian J Scholl, Objects and attention: The state of the art, Cognition 80 (1-2) (2001) 1–46.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] Tomer Ullman, Andreas Stuhlmüller, Noah Goodman, Joshua B Tenenbaum, Learning physics from dynamical scenes, in: Annual Meeting of the Cognitive Science Society (CogSci), 2014.

[74] Tobias Gerstenberg, Joshua B Tenenbaum, Intuitive theories, in: Oxford handbook of causal reasoning, Oxford University Press New York, NY, 2017, pp. 515–548.

[75] 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.

[76] Colin Maclaurin, A Treatise of Fluxions: In Two Books. 1, Vol. 1, Ruddimans, 1742.

[77] Erik T Mueller, Commonsense reasoning: an event calculus based approach, Morgan Kaufmann, 2014.

[78] Erik T Mueller, Daydreaming in humans and machines: a computer model of the stream of thought, Intellect Books, 1990.

[79] Albert Michotte, The perception of causality (TR Miles, Trans.), London, England: Methuen & Co, 1963.

[80] Susan Carey, The origin of concepts, Oxford University Press, 2009.

[81] 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.

[82] Ali Farhadi, Ian Endres, Derek Hoiem, David Forsyth, Describing objects by their attributes, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[83] Devi Parikh, Kristen Grauman, Relative attributes, in: International Conference on Computer Vision (ICCV), 2011.

[84] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, Benjamin Rozenfeld, Learning realistic human actions from movies, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[85] Benjamin Yao, Song-Chun Zhu, Learning deformable action templates from cluttered videos, in: International Conference on Computer Vi-

sion (ICCV), 2009.

[86] 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.

[87] 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.

[88] Navneet Dalal, Bill Triggs, Histograms of oriented gradients for human detection, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[89] 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.

[90] RW Fleming, M Barnett-Cowan, HH Bülthoff, Perceived object stability is affected by the internal representation of gravity, PLoS One 6 (4).

[91] 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.

[92] Philip J Kellman, Elizabeth S Spelke, Perception of partly occluded objects in infancy, Cognitive psychology 15 (4) (1983) 483–524.

[93] Renée Baillargeon, Elizabeth S Spelke, Stanley Wasserman, Object permanence in five-month-old infants, Cognition 20 (3) (1985) 191–208.

[94] Scott P Johnson, Richard N Aslin, Perception of object unity in 2-month-old infants, Developmental Psychology 31 (5) (1995) 739.

[95] Amy Needham, Factors affecting infants' use of featural information in object segregation, Current Directions in Psychological Science 6 (2) (1997) 26–33.

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

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] Gergely Csibra, György Gergely, 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans, Acta psychologica 124 (1) (2007) 60–78.

[104] Chris L Baker, Joshua B Tenenbaum, Rebecca R Saxe, Goal inference as inverse planning, in: Annual Meeting of the Cognitive Science Society (CogSci), 2007.

[105] Chris L Baker, Noah D Goodman, Joshua B Tenenbaum, Theory-based social goal inference, in: Annual Meeting of the Cognitive Science Society (CogSci), 2008.

[106] Minh Hoai, Fernando De la Torre, Max-margin early event detectors, International Journal of Computer Vision (IJCV) 107 (2) (2014) 191–202.

[107] Matthew W Turek, Anthony Hoogs, Roderic Collins, Unsupervised learning of functional categories in video scenes, in: European Conference on Computer Vision (ECCV), 2010.

[108] Helmut Grabner, Juergen Gall, Luc Van Gool, What makes a chair a chair?, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] Noam Nisan, Amir Ronen, Algorithmic mechanism design, Games and Economic behavior 35 (1-2) (2001) 166–196.

[117] Jeremy Bentham, An introduction to the principles of morals, London: Athlone.

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

[119] Brian J Scholl, Patrice D Tremoulet, Perceptual causality and animacy, Trends in Cognitive Sciences 4 (8) (2000) 299–309.

[120] Alfred Arthur Robb, Optical geometry of motion: A new view of the theory of relativity, W. Heffer, 1911.

[121] David B Malament, The class of continuous timelike curves determines the topology of spacetime, Journal of mathematical physics 18 (7) (1977) 1399–1404.

[122] Alfred A Robb, Geometry of time and space, Cambridge University Press, 2014.

[123] Roberta Corrigan, Peggy Denton, Causal understanding as a developmental primitive, Developmental review 16 (2) (1996) 162–202.

[124] Peter A White, Causal processing: Origins and development, Psychological bulletin 104 (1) (1988) 36.

[125] 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.

[126] Keith Holyoak, Patricia W. Cheng, Causal learning and inference as a rational process: The new synthesis, Annual Review of Psychology 62 (2011) 135–163.

[127] D. R. Shanks, A. Dickinson, Associative accounts of causality judgment, Psychology of learning and motivation 21 (1988) 229–261.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] Patricia W Cheng, From covariation to causation: a causal power theory, Psychological Review 104 (2) (1997) 367–405.

[134] Martin Rolfs, Michael Dambacher, Patrick Cavanagh, Visual adaptation of the perception of causality, Current Biology 23 (3) (2013) 250–254.

[135] Celeste McCollough, Color adaptation of edge-detectors in the human visual system, Science 149 (3688) (1965) 1115–1116.

[136] 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.

[137] Tobias Gerstenberg, Matthew F Peterson, Noah D Goodman, David A Lagnado, Joshua B Tenenbaum, Eye-tracking causality, Psychological Science 28 (12) (2017) 1731–1744.

[138] 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.

[139] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, Philipp Moritz, Trust region policy optimization, in: International Conference on Machine Learning (ICML), 2015.

[140] 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.

[141] 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.

[142] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347.

[143] Chiyuan Zhang, Oriol Vinyals, Remi Munos, Samy Bengio, A study on overfitting in deep reinforcement learning, arXiv preprint arXiv:1804.06893.

[144] 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.

[145] Mark Edmonds, Xiaojian Ma, Siyuan Qi, Yixin Zhu, Hongjing Lu, Song-Chun Zhu, Theory-based causal transfer: Integrating instance-level induction and abstract-level structure learning, in: AAAI, 2020.

[146] Donald B Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies., Journal of educational Psychology 66 (5) (1974) 688.

[147] Guido W Imbens, Donald B Rubin, Causal inference in statistics, social, and biomedical sciences, Cambridge University Press, 2015.

[148] 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.

[149] J Pearl, Causality: Models, reasoning and inference, Cambridge University Press, 2000.

[150] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, Thomas Richardson, Causation, prediction, and search, MIT press, 2000.

[151] David Maxwell Chickering, Optimal structure identification with greedy search, Journal of machine learning research 3 (Nov) (2002) 507–554.

[152] 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.

[153] 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.

[154] 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.

[155] Amy Fire, Song-Chun Zhu, Learning perceptual causality from video, ACM Transactions on Intelligent Systems and Technology (TIST) 7 (2) (2016) 23.

[156] Ronald Aylmer Fisher, The design of experiments, Oliver And Boyd; Edinburgh; London, 1937.

[157] 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.

[158] Song-Chun Zhu, Ying Nian Wu, David Mumford, Minimax entropy principle and its application to texture modeling, Neural computation 9 (8) (1997) 1627–1660.

[159] 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.

[160] 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.

[161] 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.

[162] 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.

[163] Andrea A DiSessa, Unlearning aristotelian physics: A study of knowledge-based learning, Cognitive science 6 (1) (1982) 37–75.

[164] Mary Kister Kaiser, John Jonides, Joanne Alexander, Intuitive reasoning about abstract and familiar physics problems, Memory & Cognition 14 (4) (1986) 308–312.

[165] Kevin A Smith, Peter Battaglia, Edward Vul, Consistent physics underlying ballistic motion prediction, in: Annual Meeting of the Cognitive Science Society (CogSci), 2013.

[166] 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.

[167] 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.

[168] 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.

[169] Jean Piaget, Margaret Cook, The origins of intelligence in children, International Universities Press New York, 1952.

[170] Jean Piaget, Margaret Trans Cook, The construction of reality in the child., Basic Books, 1954.

[171] 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.

[172] 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.

[173] T GR Bower, Development in infancy, WH Freeman, 1974.

[174] Alan M Leslie, Stephanie Keeble, Do six-month-old infants perceive causality?, Cognition 25 (3) (1987) 265–288.

[175] 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.

[176] 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.

[177] 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.

[178] Peter Achinstein, The nature of explanation, Oxford University Press on Demand, 1983.

[179] 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.

[180] 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.

[181] 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.

[182] 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.

[183] 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.

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

[185] David Mumford, Agnès Desolneux, Pattern theory: the stochastic analysis of real-world signals, AK Peters/CRC Press, 2010.

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

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

[188] 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.

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

[190] 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.

[191] 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.

[192] 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.

[193] Mark Nitzberg, David Mumford, The 2.1-d sketch, in: ICCV, 1990.

[194] John YA Wang, Edward H Adelson, Layered representation for motion analysis, in: Conference on Computer Vision and Pattern Recognition (CVPR), 1993.

[195] John YA Wang, Edward H Adelson, Representing moving images with layers, Transactions on Image Processing (TIP) 3 (5) (1994) 625–638.

[196] 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.

[197] I Binford, Visual perception by computer, in: IEEE Conference of

Systems and Control, 1971.

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

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

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

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

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

[203] Max Wertheimer, Experimentelle studien uber das sehen von bewegung [experimental studies on the seeing of motion], Zeitschrift fur Psychologie 61 (1912) 161–265.

[204] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, Rüdiger von der Heydt, A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization., Psychological bulletin 138 (6) (2012) 1172.

[205] Johan Wagemans, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R Pomerantz, Peter A Van der Helm, Cees Van Leeuwen, A century of gestalt psychology in visual perception: Ii. conceptual and theoretical foundations., Psychological bulletin 138 (6) (2012) 1218.

[206] Wolfgang Köhler, Die physischen Gestalten in Ruhe und im stationärenZustand. Eine natur-philosophische Untersuchung [The physical Gestalten at rest and in steady state], Braunschweig, Germany: Vieweg und Sohn., 1920.

[207] Wolfgang Köhler, Physical gestalten, in: A source book of Gestalt psychology, London, England: Routledge & Kegan Paul, 1938, pp. 17–54.

[208] Max Wertheimer, Untersuchungen zur lehre von der gestalt, ii. [investigations in gestalt theory: Ii. laws of organization in perceptual forms], Psychologische Forschung 4 (1923) 301–350.

[209] Max Wertheimer, Laws of organization in perceptual forms, in: A source book of Gestalt psychology, London, England: Routledge & Kegan Paul, 1938, pp. 71–94.

[210] Kurt Koffka, Principles of Gestalt psychology, Routledge, 2013.

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

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

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

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

[215] Robert L Solso, M Kimberly MacLin, Otto H MacLin, Cognitive psychology, Pearson Education New Zealand, 2005.

[216] Peter Dayan, Geoffrey E Hinton, Radford M Neal, Richard S Zemel, The helmholtz machine, Neural computation 7 (5) (1995) 889–904.

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

[218] Irving Biederman, Robert J Mezzanotte, Jan C Rabinowitz, Scene perception: Detecting and judging objects undergoing relational violations, Cognitive psychology (1982) 143–177.

[219] Manuel Blum, Arnold Griffith, Bernard Neumann, A stability test for configurations of blocks, Tech. rep., Massachusetts Institute of Technology (1970).

[220] 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.

[221] 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.

[222] Varsha Hedau, Derek Hoiem, David Forsyth, Recovering the spatial layout of cluttered rooms, in: International Conference on Computer Vision (ICCV), 2009.

[223] David C Lee, Martial Hebert, Takeo Kanade, Geometric reasoning for single image structure recovery, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[224] 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.

[225] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, Rob Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision (ECCV), 2012.

[226] 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.

[227] Ruiqi Guo, Derek Hoiem, Support surface prediction in indoor scenes, in: International Conference on Computer Vision (ICCV), 2013.

[228] 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).

[229] 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.

[230] 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.

[231] 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.

[232] 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.

[233] 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.

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

[235] Marcus A Brubaker, Leonid Sigal, David J Fleet, Estimating contact dynamics, in: International Conference on Computer Vision (ICCV), 2009.

[236] 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.

[237] 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.

[238] 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.

[239] 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.

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

[241] James Jerome Gibson, The senses considered as perceptual systems, Houghton Mifflin, 1966.

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

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

[244] Mohammed Hassanin, Salman Khan, Murat Tahtali, Visual affordance and function understanding: A survey, arXiv preprint arXiv:1807.06775.

[245] 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.

[246] Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, Data-driven grasp synthesis—a survey, IEEE Transactions on Robotics 30 (2) (2013) 289–309.

[247] 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.

[248] Wolfgang Kohler, The mentality of apes, New York: Liverright, 1925.

[249] William Homan Thorpe, Learning and instinct in animals, Harvard University Press, 1956.

[250] Kenneth Page Oakley, Man the tool-maker, University of Chicago Press, 1968.

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

[252] 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.

[253] 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.

[254] 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.

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

[256] Alex AS Weir, Jackie Chappell, Alex Kacelnik, Shaping of hooks in new caledonian crows, Science 297 (5583) (2002) 981–981.

[257] 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.

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

[259] 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.

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

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

[262] 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.

[263] 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.

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

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

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

[267] Marvin Minsky, Society of mind, Simon and Schuster, 1988.

[268] 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.

[269] Zhijian Liu, William T Freeman, Joshua B Tenenbaum, Jiajun Wu, Physical primitive decomposition, in: European Conference on Computer Vision (ECCV), 2018.

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

[271] 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.

[272] 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.

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

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

[275] 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.

[276] 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.

[277] Ernest Davis, Gary Marcus, Noah Frazier-Logue, Commonsense reasoning about containers using radically incomplete information, Artificial intelligence 248 (2017) 46–84.

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

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

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

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

[282] 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.

[283] 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.

[284] 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.

[285] 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.

[286] 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.

[287] Yun Jiang, Marcus Lim, Ashutosh Saxena, Learning object arrangements in 3d scenes using human context, in: International Conference on Machine Learning (ICML), 2012.

[288] 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.

[289] Kerstin Dautenhahn, Chrystopher L Nehaniv, Imitation in Animals and Artifacts, MIT Press Cambridge, MA, 2002.

[290] 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.

[291] 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.

[292] 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.

[293] 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.

[294] 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.

[295] 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.

[296] Mark Edmonds, Feng Gao, Hangxin Liu, Xu Xie, Siyuan Qi, Brandon Rothrock, Yixin Zhu, Ying Nian Wu, Hongjing Lu, Song-Chun Zhu, A tale of two explanations: Enhancing human trust by explaining robot behavior, Science Robotics 4 (37).

[297] 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.

[298] 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.

[299] Daniel Clement Dennett, The intentional stance, MIT press, 1989.

[300] Fritz Heider, The psychology of interpersonal relations, Psychology Press, 2013.

[301] György Gergely, Zoltán Nádasdy, Gergely Csibra, Szilvia Bíró, Taking the intentional stance at 12 months of age, Cognition 56 (2) (1995) 165–193.

[302] David Premack, Guy Woodruff, Does the chimpanzee have a theory of mind?, Behavioral and brain sciences 1 (4) (1978) 515–526.

[303] Dare A Baldwin, Jodie A Baird, Discerning intentions in dynamic human action, Trends in Cognitive Sciences 5 (4) (2001) 171–178.

[304] Amanda L Woodward, Infants selectively encode the goal object of an actor's reach, Cognition 69 (1) (1998) 1–34.

[305] 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.

[306] Dare A Baldwin, Jodie A Baird, Megan M Saylor, M Angela Clark, Infants parse dynamic action, Child development 72 (3) (2001) 708–717.

[307] 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.

[308] Szilvia Biro, Bernhard Hommel, Becoming an intentional agent: introduction to the special issue., Acta psychologica 124 (1) (2007) 1–7.

[309] György Gergely, Harold Bekkering, Ildikó Király, Developmental psy-

chology: Rational imitation in preverbal infants, Nature 415 (6873) (2002) 755.

[310] 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.

[311] M Tomasello, Developing theories of intention (1999).

[312] Paul Bloom, Intention, history, and artifact concepts, Cognition 60 (1) (1996) 1–29.

[313] Fritz Heider, Marianne Simmel, An experimental study of apparent behavior, The American journal of psychology 57 (2) (1944) 243–259.

[314] 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.

[315] John N Bassili, Temporal and spatial contingencies in the perception of social events, Journal of Personality and Social Psychology 33 (6) (1976) 680.

[316] Winand H Dittrich, Stephen EG Lea, Visual perception of intentional motion, Perception 23 (3) (1994) 253–268.

[317] 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.

[318] 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.

[319] Daniel C Dennett, Précis of the intentional stance, Behavioral and brain sciences 11 (3) (1988) 495–505.

[320] 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.

[321] Shari Liu, Elizabeth S Spelke, Six-month-old infants expect agents to minimize the cost of their actions, Cognition 160 (2017) 35–42.

[322] György Gergely, Gergely Csibra, Teleological reasoning in infancy: The naıve theory of rational action, Trends in Cognitive Sciences 7 (7) (2003) 287–292.

[323] Chris L Baker, Rebecca Saxe, Joshua B Tenenbaum, Action understanding as inverse planning, Cognition 113 (3) (2009) 329–349.

[324] Luís Moniz Pereira, et al., Intention recognition via causal bayes networks plus plan generation, in: Portuguese Conference on Artificial Intelligence, Springer, 2009.

[325] 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.

[326] 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.

[327] Yu Kong, Yun Fu, Human action recognition and prediction: A survey, arXiv preprint arXiv:1806.11230.

[328] Sarah-Jayne Blakemore, Jean Decety, From the perception of action to the understanding of intention, Nature reviews neuroscience 2 (8) (2001) 561.

[329] Birgit Elsner, Bernhard Hommel, Effect anticipation and action control., Journal of experimental psychology: human perception and performance 27 (1) (2001) 229.

[330] Birgit Elsner, Infants' imitation of goal-directed actions: The role of movements and action effects, Acta psychologica 124 (1) (2007) 44–59.

[331] Giacomo Rizzolatti, Laila Craighero, The mirror-neuron system, Annual Review of Neuroscience 27 (2004) 169–192.

[332] 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.

[333] 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.

[334] 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.

[335] Gergely Csibra, György Gergely, The teleological origins of mentalistic action explanations: A developmental hypothesis, Developmental Science 1 (2) (1998) 255–259.

[336] György Gergely, The development of understanding self and agency, Blackwell handbook of childhood cognitive development (2002) 26–46.

[337] Chris L Kleinke, Gaze and eye contact: a research review, Psychological bulletin 100 (1) (1986) 78.

[338] Nathan J Emery, The eyes have it: the neuroethology, function and evolution of social gaze, Neuroscience & Biobehavioral Reviews 24 (6) (2000) 581–604.

[339] Judee K Burgoon, Laura K Guerrero, Kory Floyd, Nonverbal communication, Routledge, 2016.

[340] 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.

[341] 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.

[342] 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.

[343] Susanne Trick, Dorothea Koert, Jan Peters, Constantin Rothkopf, Multimodal uncertainty reduction for intention recognition in human-robot interaction, arXiv preprint arXiv:1907.02426.

[344] Tianmin Shu, Michael S Ryoo, Song-Chun Zhu, Learning social affordance for human-robot interaction, in: International Joint Conference on Artificial Intelligence (IJCAI), 2016.

[345] 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.

[346] Stuart J Russell, Peter Norvig, Artificial intelligence: a modern approach, Malaysia; Pearson Education Limited,, 2016.

[347] Francis Hutcheson, An Inquiry into the Original of our Ideas of Beauty and Virtue: in two treatises, J. Darby...[and 8 others], 1726.

[348] John Stuart Mill, Utilitarianism, Longmans, Green and Company, 1863.

[349] 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.

[350] 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.

[351] H Paul Grice, Peter Cole, Jerry Morgan, et al., Logic and conversation, 1975 (1975) 41–58.

[352] Noah D Goodman, Michael C Frank, Pragmatic language interpretation as probabilistic inference, Trends in Cognitive Sciences 20 (11) (2016) 818–829.

[353] David Lewis, Convention: A philosophical study, John Wiley & Sons, 2008.

[354] Dan Sperber, Deirdre Wilson, Relevance: Communication and cognition, Vol. 142, Harvard University Press Cambridge, MA, 1986.

[355] Ludwig Wittgenstein, Philosophical Investigations, Macmillan, 1953.

[356] Herbert H Clark, Using language, Cambridge university press, 1996.

[357] 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.

[358] Noah D Goodman, Andreas Stuhlmüller, Knowledge and implicature: Modeling language understanding as social cognition, Topics in cognitive science 5 (1) (2013) 173–184.

[359] Robert Dale, Ehud Reiter, Computational interpretations of the gricean maxims in the generation of referring expressions, Cognitive science 19 (2) (1995) 233–263.

[360] Anton Benz, Gerhard Jäger, Robert Van Rooij, An introduction to game theory for linguists, in: Game theory and pragmatics, Springer, 2006, pp. 1–82.

[361] Gerhard Jäger, Applications of game theory in linguistics, Language and Linguistics compass 2 (3) (2008) 406–421.

[362] Michael C Frank, Noah D Goodman, Predicting pragmatic reasoning in language games, Science 336 (6084) (2012) 998–998.

[363] 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.

[364] Max Kleiman-Weiner, Mark K Ho, Joseph L Austerweil, Michael L Littman, Joshua B Tenenbaum, Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction, in: Annual Meeting of the Cognitive Science Society (CogSci), 2016.

[365] Michael Shum, Max Kleiman-Weiner, Michael L Littman, Joshua B Tenenbaum, Theory of minds: Understanding behavior in groups through inverse planning, in: AAAI Conference on Artificial Intelligence (AAAI), 2019.

[366] Max Kleiman-Weiner, Alex Shaw, Josh Tenenbaum, Constructing social preferences from anticipated judgments: When impartial inequity is fair and why?, in: Annual Meeting of the Cognitive Science Society (CogSci), 2017.

[367] Max Kleiman-Weiner, Rebecca Saxe, Joshua B Tenenbaum, Learning a commonsense moral theory, cognition 167 (2017) 107–123.

[368] Michael Kinney, Costas Tsatsoulis, Learning communication strategies in multiagent systems, Applied intelligence 9 (1) (1998) 71–91.

[369] 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.

[370] 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.

[371] 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.

[372] Keith J Holyoak, Analogy and relational reasoning, in: The Oxford Handbook of Thinking and Reasoning, Oxford University Press, 2012, pp. 234–259.

[373] J. C. et al. Raven, Raven's progressive matrices, Western Psychological Services.

[374] 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.

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

[376] 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.

[377] 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.

[378] 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.

[379] 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.

[380] 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.

[381] 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.

[382] Yi Wu, Yuxin Wu, Georgia Gkioxari, Yuandong Tian, Building generalizable agents with a realistic and rich 3d environment, arXiv preprint arXiv:1801.02209.

[383] 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.

[384] 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.

[385] 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.

[386] 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.

[387] 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).

[388] Demetri Terzopoulos, John Platt, Alan Barr, Kurt Fleischer, Elastically deformable models, ACM Transactions on Graphics (TOG) 21 (4) (1987) 205–214.

[389] Demetri Terzopoulos, Kurt Fleischer, Modeling inelastic deformation: viscolelasticity, plasticity, fracture, ACM Transactions on Graphics (TOG) 22 (4) (1988) 269–278.

[390] Nick Foster, Dimitri Metaxas, Realistic animation of liquids, Graphical models and image processing 58 (5) (1996) 471–483.

[391] Jos Stam, Stable fluids, in: ACM Transactions on Graphics (TOG), Vol. 99, 1999.

[392] Robert Bridson, Fluid simulation for computer graphics, CRC Press, 2015.

[393] Javier Bonet, Richard D Wood, Nonlinear continuum mechanics for finite element analysis, Cambridge university press, 1997.

[394] 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.

[395] 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.

[396] 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.

[397] 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.

[398] 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.

[399] J. J. Monaghan, Smoothed particle hydrodynamics, Annual review of astronomy and astrophysics 30 (1) (1992) 543–574.

[400] 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.

[401] S. Li, W. K. Liu, Meshfree and particle methods and their applications, Applied Mechanics Reviews 55 (1) (2002) 1–34.

[402] 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.

[403] 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.

[404] 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.

[405] 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.

[406] 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.

[407] 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.

[408] Johan Gaume, T Gast, J Teran, A van Herwijnen, C Jiang, Dynamic anticrack propagation in snow, Nature communications 9 (1) (2018) 3047.

[409] 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.

[410] 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.

[411] 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.

[412] 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.

[413] 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.

[414] 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.

[415] 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.

[416] 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.

[417] Chenfanfu Jiang, Theodore Gast, Joseph Teran, Anisotropic elastoplasticity for cloth, knit and hair frictional contact, ACM Transactions on Graphics (TOG) 36 (4) (2017) 152.

[418] Xuchen Han, Theodore F Gast, Qi Guo, Stephanie Wang, Chenfanfu Jiang, Joseph Teran, A hybrid material point method for frictional con-

tact with diverse materials, ACM Transactions on Graphics (TOG) 2 (2) (2019) 17.

[419] 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.

[420] 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.

[421] 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.

[422] 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.

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

[424] 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.

[425] 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.

[426] 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.

[427] 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.

[428] 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.

[429] 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.

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

[431] 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.

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

[433] Yu Fang, Yuanming Hu, Shi-Min Hu, Chenfanfu Jiang, A temporally adaptive material point method with regional time stepping, in: Computer Graphics Forum, 2018.

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

[435] 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.

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

[437] 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.

[438] 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.

[439] Claudia V Goldman, Shlomo Zilberstein, Optimizing information exchange in cooperative multi-agent systems, in: International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2003.

[440] 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.

[441] 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.

[442] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, Shimon Whiteson, Counterfactual multi-agent policy gradients, in: AAAI Conference on Artificial Intelligence (AAAI), 2018.

[443] Sainbayar Sukhbaatar, Rob Fergus, et al., Learning multiagent communication with backpropagation, in: Advances in Neural Information Processing Systems (NeurIPS), 2016.

[444] Igor Mordatch, Pieter Abbeel, Emergence of grounded compositional language in multi-agent populations, in: AAAI Conference on Artificial Intelligence (AAAI), 2018.

[445] Angeliki Lazaridou, Alexander Peysakhovich, Marco Baroni, Multi-agent cooperation and the emergence of (natural) language, in: International Conference on Learning Representations (ICLR), 2017.

[446] 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.

[447] Katrina Evtimova, Andrew Drozdov, Douwe Kiela, Kyunghyun Cho, Emergent language in a multi-modal, multi-step referential game, arXiv preprint arXiv:1705.10369.

[448] 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.

[449] 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.

[450] 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.

[451] Laura Graesser, Kyunghyun Cho, Douwe Kiela, Emergent linguistic phenomena in multi-agent communication games, arXiv preprint arXiv:1901.08706.

[452] Emmanuel Dupoux, Pierre Jacob, Universal moral grammar: a critical appraisal, Trends in Cognitive Sciences 11 (9) (2007) 373–378.

[453] John Mikhail, Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment, Cambridge University Press, 2011.

[454] 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.

[455] 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.

[456] 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.

[457] 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.

[458] Thomas Hurka, Virtue, vice, and value, Oxford University Press, 2000.

[459] John Rawls, A theory of justice, Harvard university press, 1971.

[460] Jonathan Haidt, The new synthesis in moral psychology, science 316 (5827) (2007) 998–1002.

[461] 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.

[462] 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.

[463] Keith J Holyoak, Paul Thagard, The analogical mind, American psychologist 52 (1) (1997) 35.

[464] Patricia W Cheng, Marc J Buehner, Causal learning, in: The Oxford Handbook of Thinking and Reasoning, Oxford University Press, 2012, pp. 210–233.

[465] Mary B Hesse, Models and analogies in science, Notre Dame University Press, 1966.

[466] 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.

[467] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[468] 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.

[469] 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.

[470] R E Snow, Patrick Kyllonen, B Marshalek, The topography of ability and learning correlations, Advances in the psychology of human intelligence (1984) 47–103.

[471] 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.

[472] Gordon H Bower, A contrast effect in differential conditioning, Journal of Experimental Psychology 62 (2) (1961) 196.

[473] Donald R Meyer, The effects of differential rewards on discrimination reversal learning by monkeys, Journal of Experimental Psychology 41 (4) (1951) 268.

[474] 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.

[475] 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.

[476] 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.

[477] Abram Amsel, Frustrative nonreward in partial reinforcement and discrimination learning: Some recent history and a theoretical extension, Psychological review 69 (4) (1962) 306.

[478] James J Gibson, Eleanor J Gibson, Perceptual learning: Differentiation or enrichment?, Psychological review 62 (1) (1955) 32.

[479] James J Gibson, The ecological approach to visual perception: classic edition, Psychology Press, 2014.

[480] Richard Catrambone, Keith J Holyoak, Overcoming contextual limitations on problem-solving transfer, Journal of Experimental Psychology: Learning, Memory, and Cognition 15 (6) (1989) 1147.

[481] Dedre Gentner, Virginia Gunn, Structural alignment facilitates the noticing of differences, Memory & Cognition 29 (4) (2001) 565–577.

[482] Rubi Hammer, Gil Diesendruck, Daphna Weinshall, Shaul Hochstein, The development of category learning strategies: What makes the difference?, Cognition 112 (1) (2009) 105–119.

[483] 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.

[484] Etsuko Haryu, Mutsumi Imai, Hiroyuki Okada, Object similarity bootstraps young children to action-based verb extension, Child Development 82 (2) (2011) 674–686.

[485] Linsey Smith, Dedre Gentner, The role of difference-detection in learning contrastive categories, in: Annual Meeting of the Cognitive Science Society (CogSci), 2014.

[486] Dedre Gentner, Structure-mapping: A theoretical framework for analogy, Cognitive science 7 (2) (1983) 155–170.

[487] Dedre Gentner, Arthur B Markman, Structural alignment in comparison: No difference without similarity, Psychological science 5 (3) (1994) 152–158.

[488] 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.

[489] 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.

[490] Stanislas Dehaene, The number sense: How the mind creates mathematics, OUP USA, 2011.

[491] Wenhe Zhang, Chi Zhang, Yixin Zhu, Song-Chun Zhu, Machine number sense: A dataset of visual arithmetic problems for abstract and relational reasoning, in: AAAI Conference on Artificial Intelligence (AAAI), 2020.