Engineering 6 (2020) 310–345

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Feature Article

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, University of California, Los Angeles, CA 90095, USA
[b] Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 16 September 2019
Revised 11 December 2019
Accepted 3 January 2020
Available online 22 February 2020

Keywords:
Computer vision
Artificial intelligence
Causality
Intuitive physics
Functionality
Perceived intent
Utility

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.

1. A call for a paradigm shift in vision and artificial intelligence

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: ① the ventral pathway for categorical recognition of objects and scenes, and ② 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.

* Corresponding author.
E-mail address: yixin.zhu@ucla.edu (Y. Zhu).

https://doi.org/10.1016/j.eng.2020.01.011

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%).[†]

The properties and characteristics of dark matter and dark energy cannot be observed and must be reasoned from 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 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 down-stream 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 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,

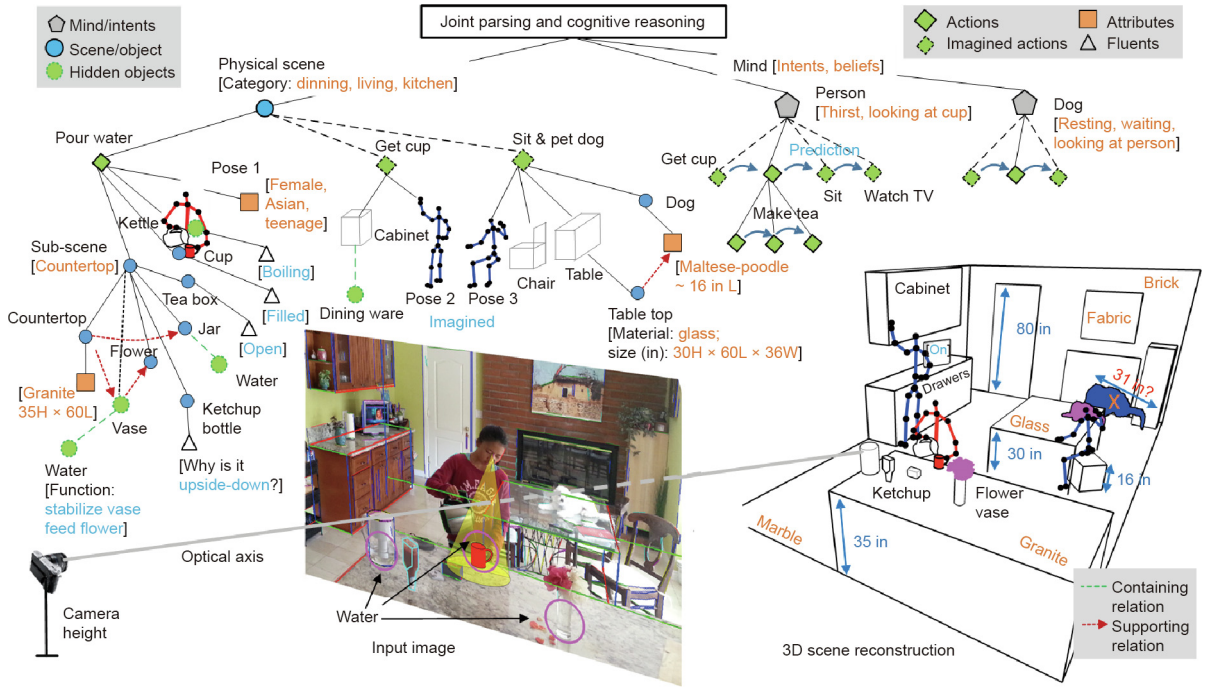

Fig. 1. An example of in-depth understanding of a scene or event through joint parsing and cognitive reasoning. From a single image, a computer vision system should be able to jointly ① reconstruct the 3D scene; ② estimate camera parameters, materials, and illumination; ③ parse the scene hierarchically with attributes, fluents, and relationships; ④ reason about the intentions and beliefs of agents (e.g., the human and dog in this example); ⑤ predict their actions in time; and ⑥ recover invisible elements such as water, latent object states, and so forth. We, as humans, can effortlessly ① predict that water is about to come out of the kettle; ② reason that the intent behind putting the ketchup bottle upside down is to utilize gravity for easy use; and ③ 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.

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

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: functionality, physics, intent, causality, and utility (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 [3]. In the kitchen shown in Fig. 2 [4], 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 [4].

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.

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 [7,8], and by Schyns and Oliva [9] and Thorpe et al. [10] in the 1990s. This line of work often leads researchers to treat categorization as a data-driven process [11–15], mostly in a feed-forward network

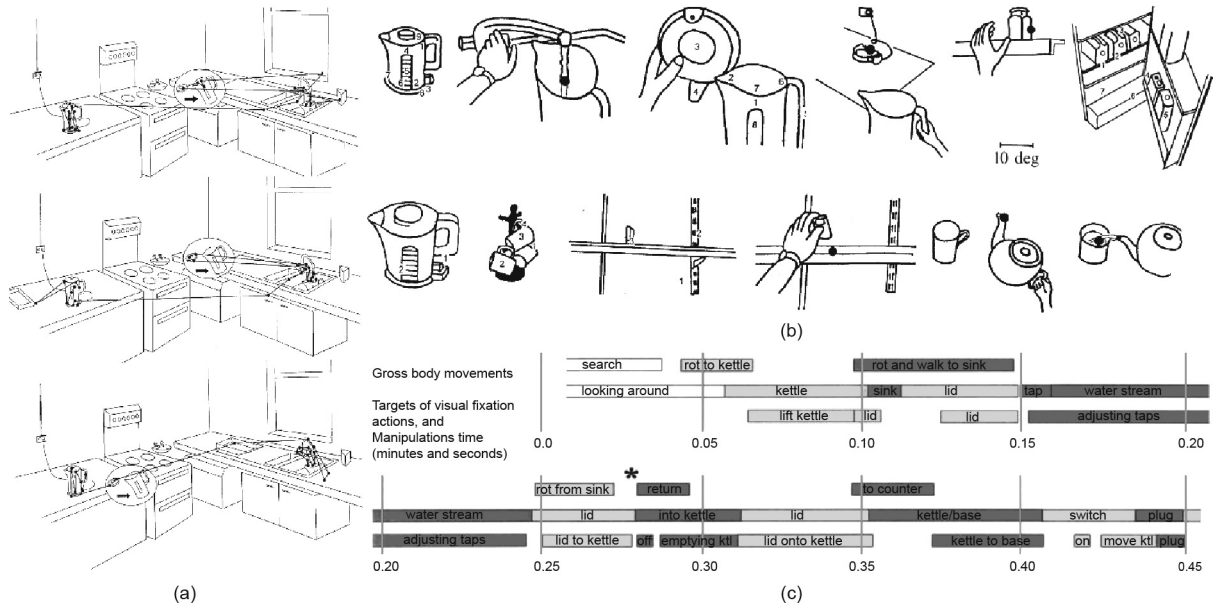

Fig. 2. Even for as "simple" a task as making a cup of tea, a person can make use of his or her single vision system to perform a variety of subtasks in order to achieve the ultimate goal. (a) Record of the visual fixations of three different subjects performing the same task of making a cup of tea in 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. [4] with permission of SAGE Publication, ©1999.

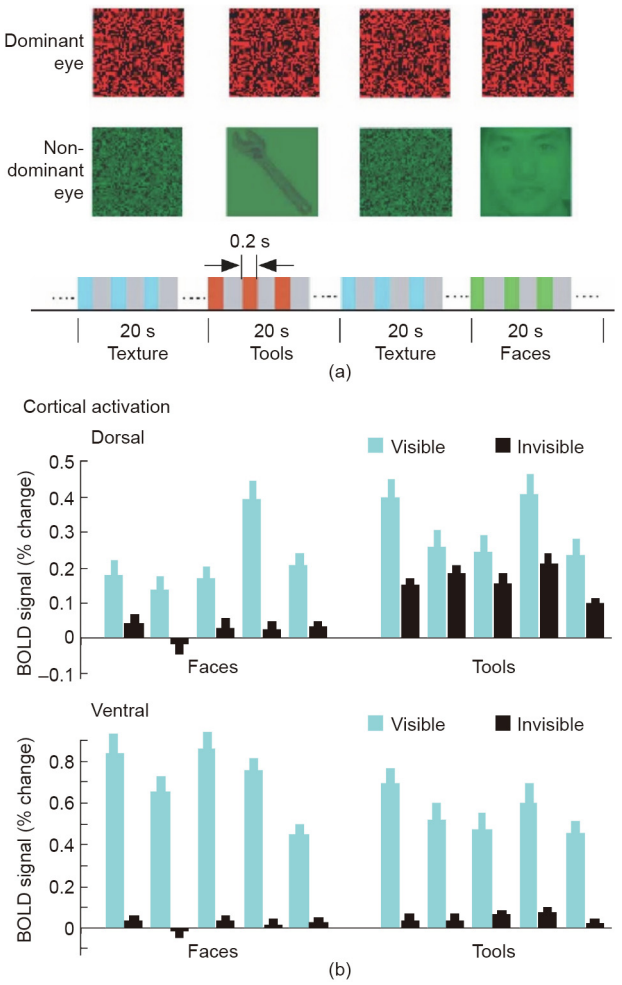

Fig. 3. Cortical responses to invisible objects in the human dorsal and ventral pathways. (a) Stimuli (tools and faces) and experimental procedures; (b) both the dorsal and ventral areas responded to tools and faces. When stimuli were suppressed by high-contrast dynamic textures, the dorsal response 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.

architecture [16,17]. 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 [18–20].

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 [21] 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 [22,23]. 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. 4) [24]. As shown in Ref. [24], scene categorization and the information-gathering process are constrained by these categorization tasks [25,26], suggesting a bidirectional interplay between the visual input and the viewer's needs/tasks [23]. Beyond scene categorization, similar phenomena were also observed in facial recognition [27].

In an early work, Ikeuchi and Hebert [28] 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. 5) [28]. 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 [31–38]; 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 [42–47]; human vision is error-prone and distorted in terms of localization [48–52]. 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 [54–56]. 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

Fig. 4. The experiment presented in Ref. [24], 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. [24] with permission of the authors, ©2014.

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)

Fig. 5. Different grasping strategies require various functional capabilities. Reproduced from Ref. [28] with permission of IEEE, ©1992.

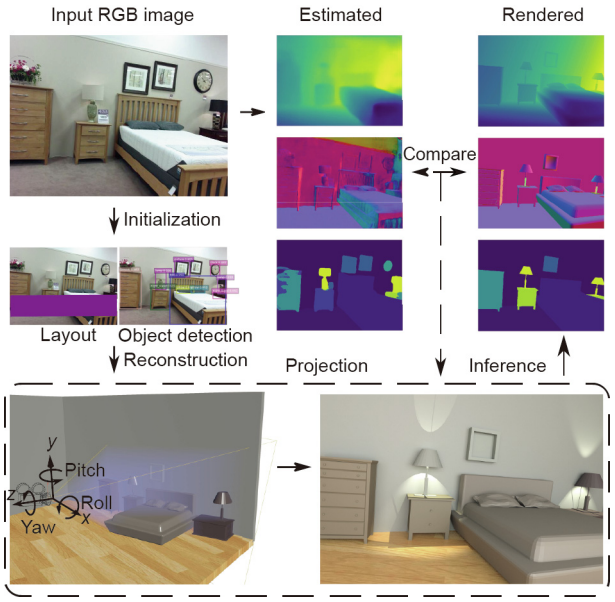

Fig. 6. Illustration of 3D indoor scene parsing and reconstruction in an analysis-by-synthesis fashion [36]. A 3D representation is initialized by individual vision tasks (e.g., object detection, 2D layout estimation). A joint inference algorithm compares the differences between the rendered normal,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

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.

2.3. Beyond "what" and "where": Toward 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) [67–69], 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. 7.

To enable an artificial agent with similar capabilities, we call for joint reasoning algorithms on a joint representation that integrates ① the "visible" traditional recognition and categorization of objects, scenes, actions, events, and so forth; and ② 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,

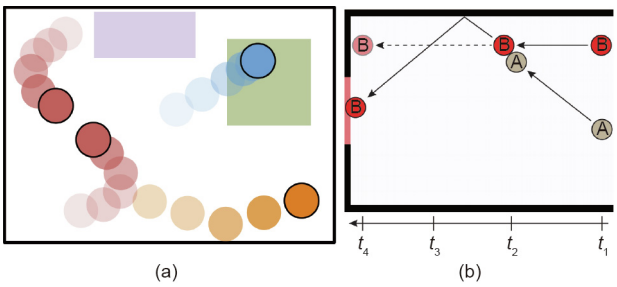

Fig. 7. (a) An animation illustrates the intent, mood, and role of the agents [73]. The motion and interaction of four different pucks moving on a 2D plane are governed by latent physical properties and dynamic laws such as mass, friction, and global and pairwise forces. (b) Intuitive theory and counterfactual 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.

a door being locked, a car blinking to signal a left turn, and a telephone ringing; see Fig. 8 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].

Fluents and perceived causality are different from the visual attributes [81,82] 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 pose–object interactions such as drinking and smoking [83–86]; 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 [87,88]. 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.

2.3.2. Intuitive physics

Psychology studies suggest that approximate Newtonian principles underlie human judgments about dynamics and stability [89,90]. 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 [91–95].

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 [96–98], 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 (Fig. 9) [36,96–100]. This observation sets useful constraints to scene understanding and is important for robotics applications [96]. 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 [34,101], 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].

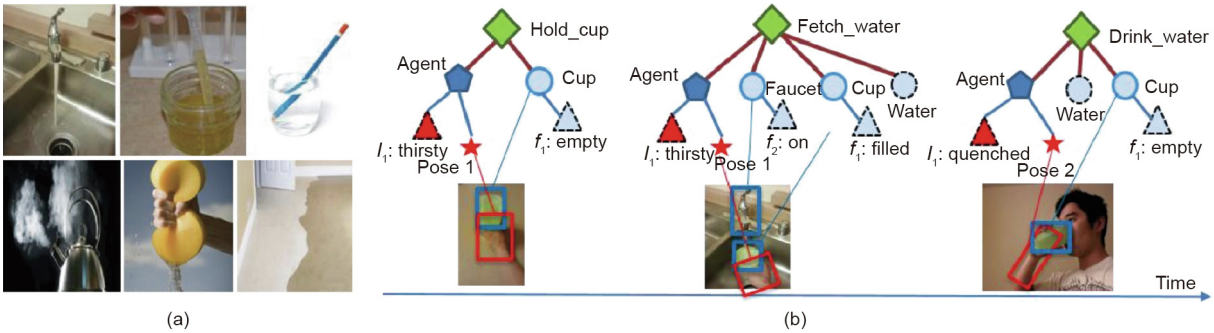

Fig. 8. Water and other clear fluids play important roles in a human's daily life, but are barely detectable in images. (a) Water causes only minor changes in appearance; (b) the "dark" entities of water, fluents (here, a cup and faucet, represented by triangles), and the intention of a human are shown in dashed nodes. The actions (diamonds) involve agents (pentagons) and cups (objects in circles).

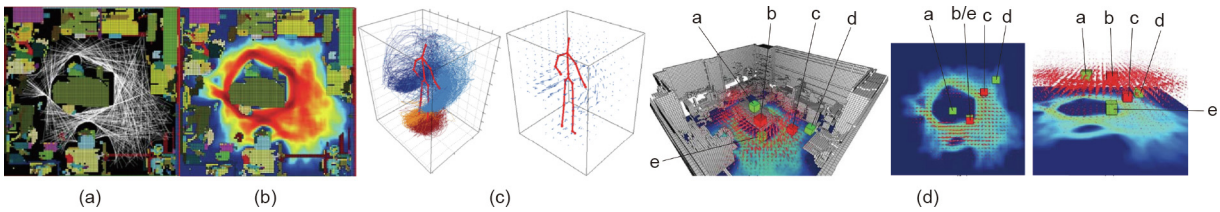

Fig. 9. Inferring the potential for objects to fall from human actions and natural disturbances. (a) The imagined human trajectories; (b) the distribution of primary motion space; (c) the secondary motion field; (d) the integrated human action field, built by integrating primary motions with secondary motions. The five objects a–e are typically a 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. [96] with permission of IEEE, ©2014.

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: ① They are hierarchically organized in a sequence of goals and are the main factors driving actions and events in a scene. ② They are completely "dark," that is, not represented by pixels. ③ Unlike the instant change of fluents in response to actions, intentions are often formed across long spatiotemporal ranges. For example, in Fig. 10 [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: ① early event and trajectory prediction [106]; ② discovery of the invisible attractive/repulsive fields of objects and recognizing their functions by analyzing human trajectories [72]; ③ understanding of scenes by function and activity [25], where the attraction fields are longer range in a scene than the functionality maps [27,107] and affordance maps [108–110] studied in recent literature; ④ understanding multi-faceted relationships among a group of people and their functional roles [111–113]; and ⑤ understanding and inferring the mental states of agents [114,115].

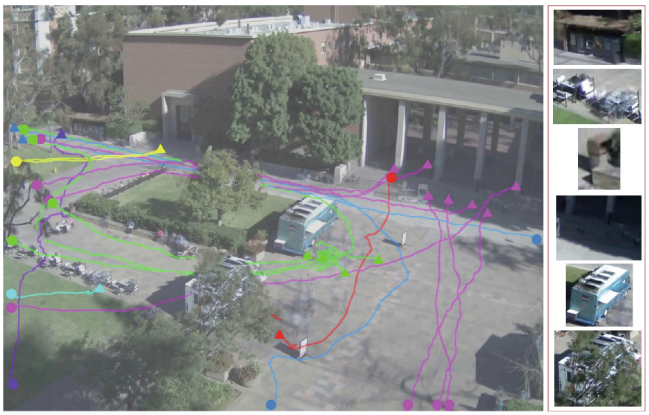

Fig. 10. People's trajectories are color-coded to indicate their shared destination. The triangles denote destinations, and the dots denote start positions; e.g., people may be heading toward the food truck to buy food (green), or to the vending machine to quench thirst (blue). Due to low resolution, 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: ① The five FPICU domains usually do not easily project onto explicit visual features; ② most of the existing computer vision and AI algorithms are neither competent in these domains nor (in most cases) applicable at all; and ③ 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 [119–121]. People have innate assumptions about causes, and causal reasoning can be activated almost automatically and irresistibly [122,123]. 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 [124] 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 way." Inherently, causal events cannot be directly interpreted merely from vision; they must be interpreted by an agent that understands the distal world [125].

Early psychological work focused on an associative mechanism as the basis for human causal learning and reasoning [126]. During this time, the Rescorla–Wagner model was used to explain how humans (and animals) build expectations using the co-occurrence of perceptual stimuli [127]. However, more recent studies have shown that human causal learning is a rational Bayesian process [125,128,129] involving the acquisition of abstract causal structure [130,131] and strength values for cause-effect relationships [132].

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. 11 [133] 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. [133] for a more detailed review.

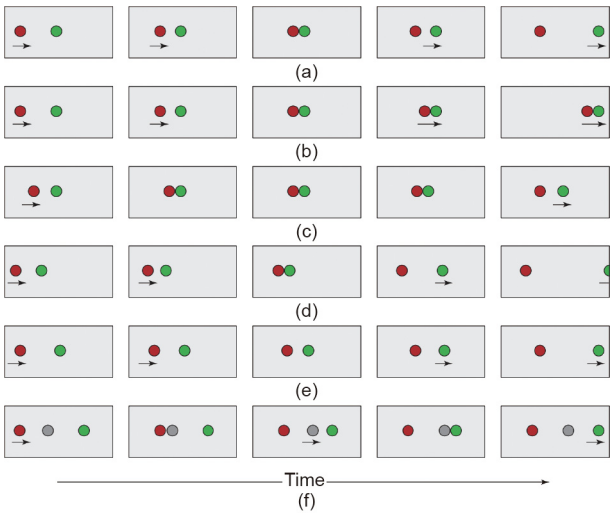

Fig. 11. Examples of some of Michotte's basic demonstrations of perceptual causality, regarding the perception of two objects, A and B (here shown as red and green circles, respectively). (a) The launching effect; (b) the entraining effect, wherein A seems to carry B along with it; (c) the launching 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. [133] with permission of Elsevier Science Ltd., ©2000.

(1) **Irresistibility.** Even if a person is told explicitly that A and B are just patches of pixels that are incapable of mechanical interactions, the person is still compelled to perceive launching. One cannot stop seeing salient causality, just as it is not possible to stop seeing color and depth.

(2) **Tightly controlled 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 than (vs. the same as) 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 [138–142], 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. [131] 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 [131]. 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.

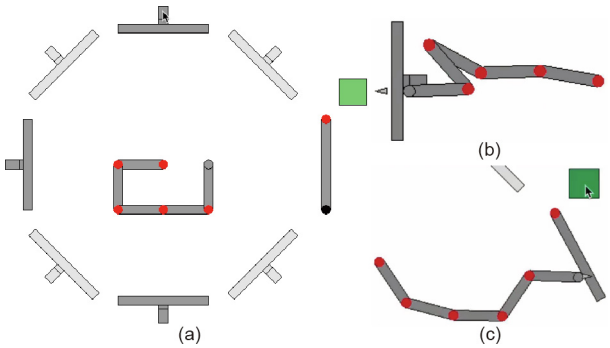

Fig. 12. The OpenLock task presented in Ref. [131]. (a) Starting configuration of a three-lever trial. All levers are being pulled toward the robot arm, whose base is anchored to the center of the display. The arm interacts with levers by either pushing outward or pulling inward. This is achieved by 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 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 constant 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: ① transferring agents between rooms tests whether or not agents form abstract representations of the environment; and ② 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 [131,145]. 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 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

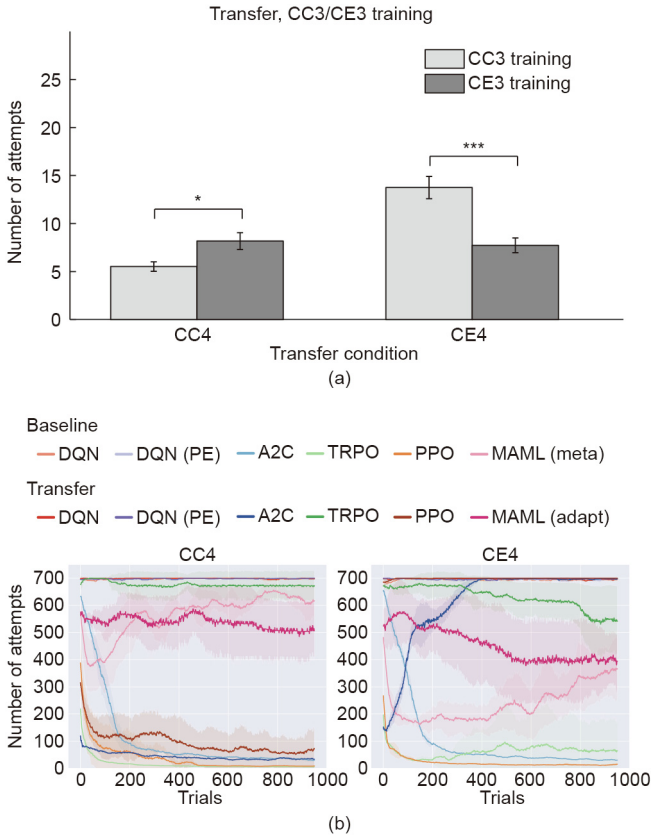

Fig. 13. Comparisons between human causal learners and typical RL agents [145]. Common Cause 4 (CC4) and Common Effect 4 (CE4) denote two transfer conditions used by Edmonds et al. [131]. (a) Average number of attempts human participants needed to find all unique solutions under four-lever Common Cause (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) and 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.

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 [150–152].

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.

3.4. Causality in computer vision

The classical and scientific clinical setting for learning causality is Fisher's randomized controlled experiments [155]. 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 [126,127,133]. 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 [156,157] proposed a method to learn "dark" causal relationships from image and video inputs, as illustrated in Fig. 14 [156]; 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

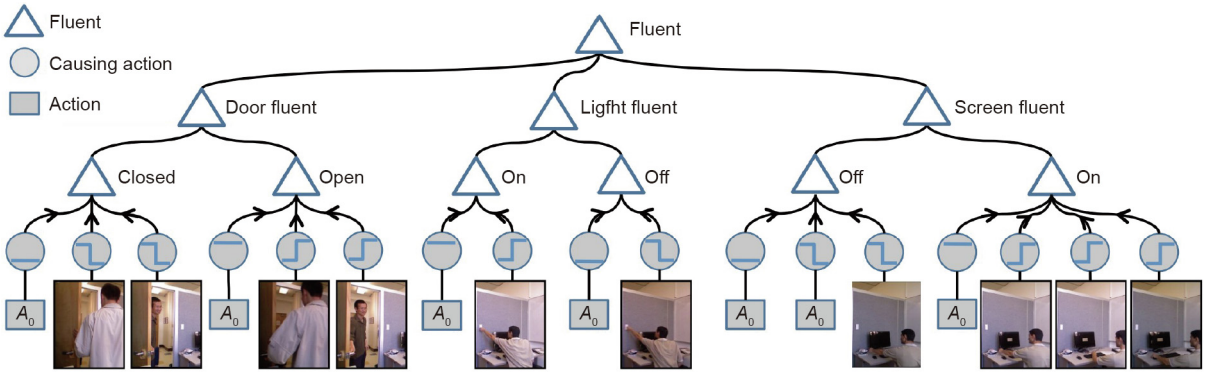

Fig. 14. An example of perceptual causality in computer vision [156], with a causal and-or graph for door status, light status, and screen status. Action A_0 represents non-action (a lack of state-changing agent action). Non-action is also used to explain the change of the monitor status to off when the screensaver activates. Arrows point from causes to effects, and undirected lines show deterministic definition.

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.

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 [162,163]. 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 [164–168].

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 [92,173–175] 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 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: ① Physical judgment is achieved by running a coarse and rough forward physical simulation; and ② 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 ① an early vision system for perceiving textures [187,188] and textons [189,190] to form a primal sketch [191,192]; ② a mid-level vision system to form 2.1D [193–195] and 2.5D [196] sketches; and ③ a high-level vision system in charge of full 3D scene formation [197–199]. In particular, Marr highlighted the importance of different levels of organization and the internal representation [200].

Alternatively, perceptual organization [201,202] and Gestalt laws [203–210] aim to resolve the 3D reconstruction problem from a single red-green-blue (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 [87,214].

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 environments; see examples in Fig. 16 [37], where systems perceive which objects

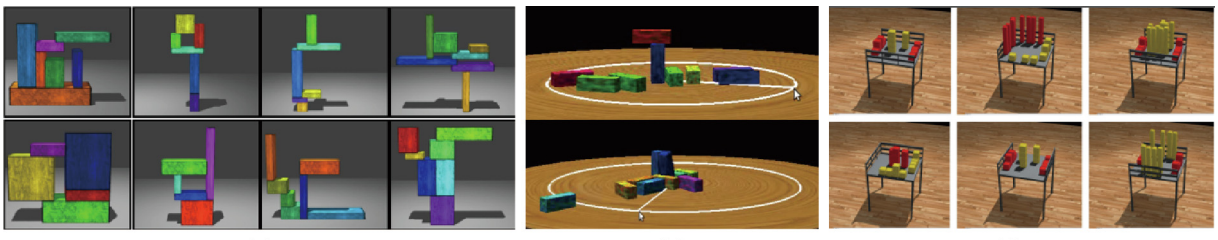

Fig. 15. Sample tasks of dynamic scene inferences about physics, stability, and support relationships presented in Ref. [70]: (a) Will it fall? (b) In which direction? (c) Which is more likely to fall if the table was bumped hard enough, the yellow or the red? Across a variety of tasks, the intuitive 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.

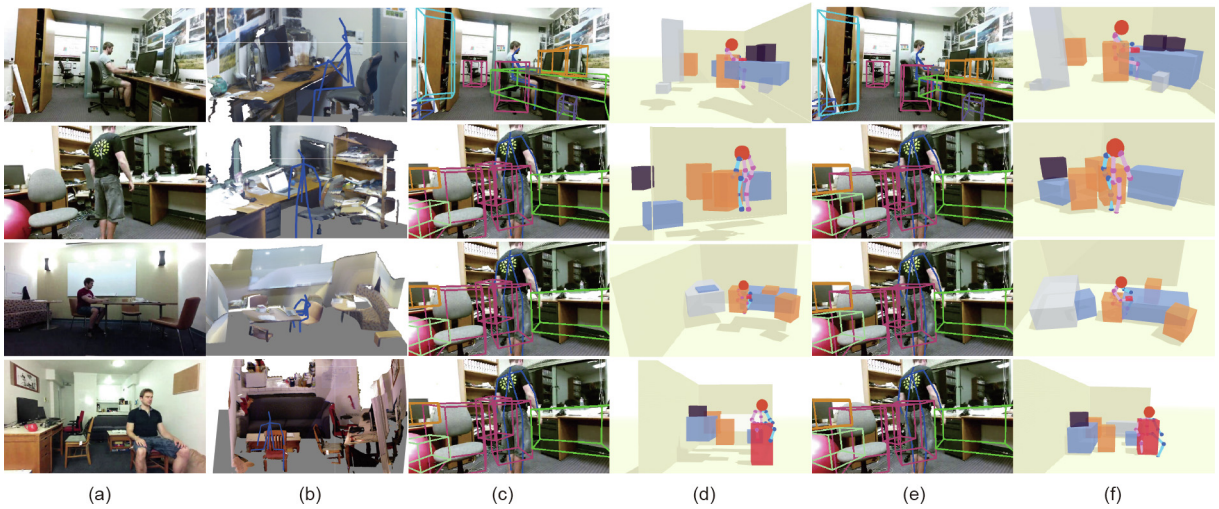

Fig. 16. Scene parsing and reconstruction by integrating physics and human–object interactions. (a) Input image; (b) ground truth; (c, d) without incorporating physics, the objects might appear to float in the air, resulting in an incorrect parsing; (e, f) after incorporating physics, the parsed 3D scene 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.

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 [32,34,109,222–228] 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. 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 [36], and enforced physical stability and plausibility by penalizing the intersections among reconstructed 3D objects and room layout [37,100]. 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 2013–2015 [96–98]. 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 [230], 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 red-green-blue-depth (RGB-D) image sequence, Zhu et al. [232] built a system that calculated various physical concepts from just a

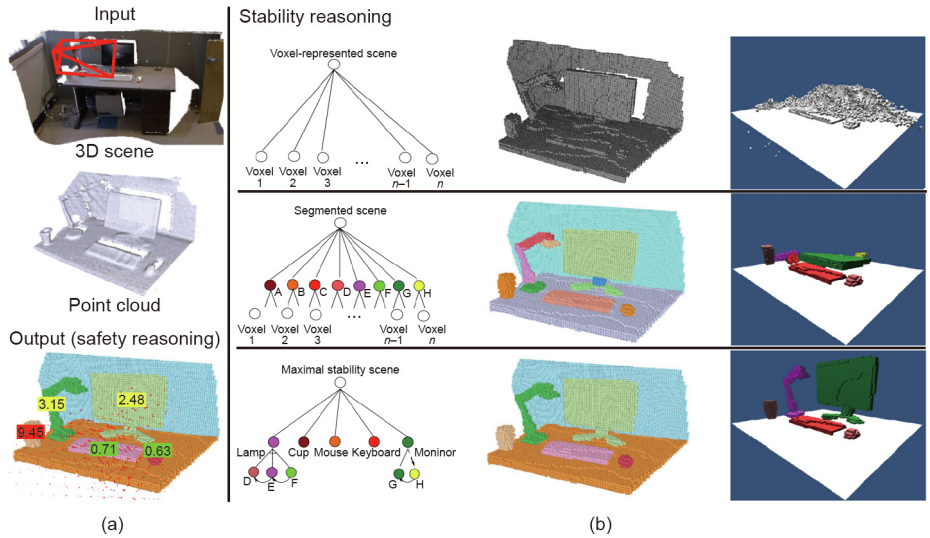

Fig. 17. An example explicitly exploiting safety and stability in a 3D scene-understanding task. Good performance in this task means that the system can understand the "dark" aspects of the image, which include how likely each object is to fall, and where the likely cause of falling will come from. (a) 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.

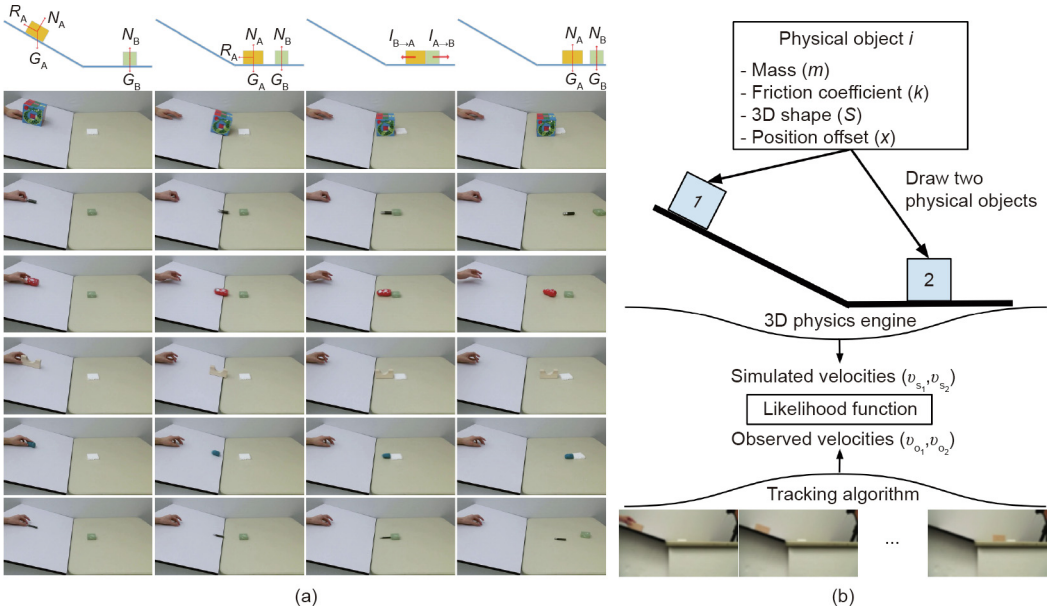

Fig. 18. Inferring the dynamics of the scenes. (a) Snapshots of the dataset; (b) overview of the Galileo model that estimates the physical properties of objects from visual inputs by incorporating the feedback of a physics engine in the loop. Reproduced from Ref. [230] with permission of Neural Information Processing Systems Foundation, Inc., ©2015.

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 (as shown in the figure in Section 7).

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. [234–236] 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

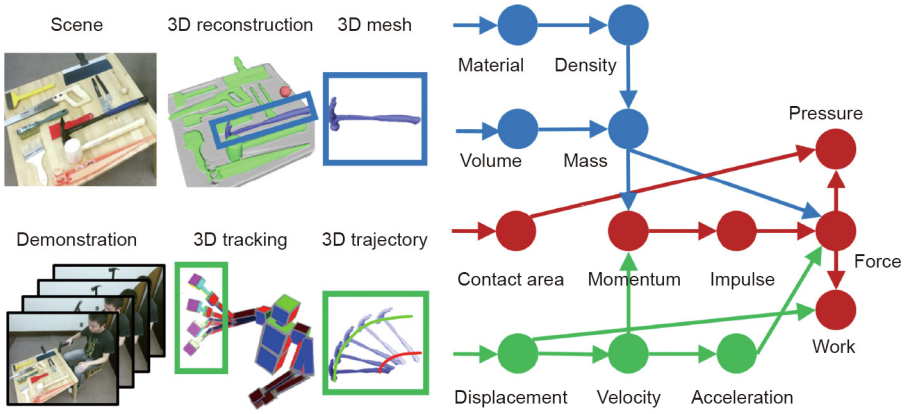

Fig. 19. Thirteen physical concepts involved in tool use and their compositional relationships. By parsing a human demonstration, the physical concepts of material, volume, concept area, and displacement are estimated from 3D meshes of tool attributes (blue), trajectories of tool use (green), or both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

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]. Section 3 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. 20. 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

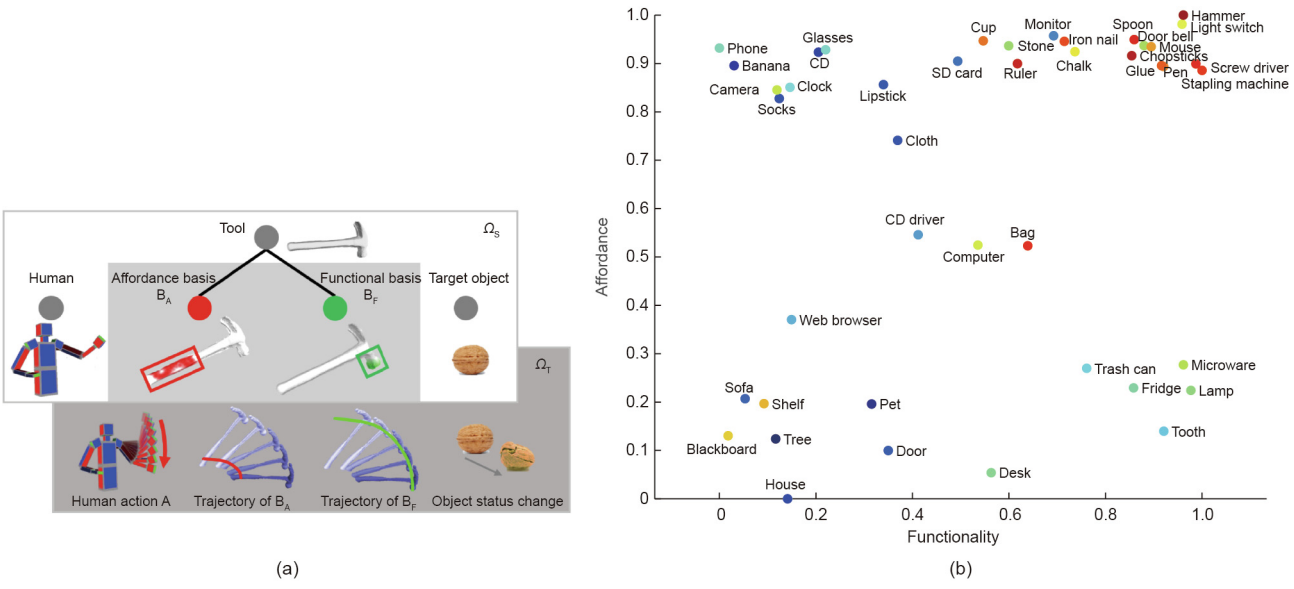

Fig. 20. (a) The task-oriented representation of a hammer and its use in cracking a nut in a joint spatiotemporal space. In this example, an object is decomposed into a functional basis and an affordance basis for a given task. (b) The likelihood of a common object being used as a tool based on its functionality 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?"

hallmark of human intelligence [250] until relatively recently, when Dr. Jane Goodall observed wild chimpanzees manufacturing and using tools with regularity [251–253]. 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.

First, why can some species devise innovative solutions, while others facing the same situation cannot? Look at the example in Fig. 21 [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.

Fig. 21. 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.

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 features 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.* (Gibson, 1977 [243])

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.

"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 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]. As early as two and a half months old, infants can already understand containers and containment [272–274]. 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 [275] and support relationships [276]). In the AI community, researchers have been adopting commonsense reasoning [277–279] and qualitative representation [280,281] for reasoning about container and containment relationships, mostly focusing on ontology, topology, first-order logic, and knowledge base.

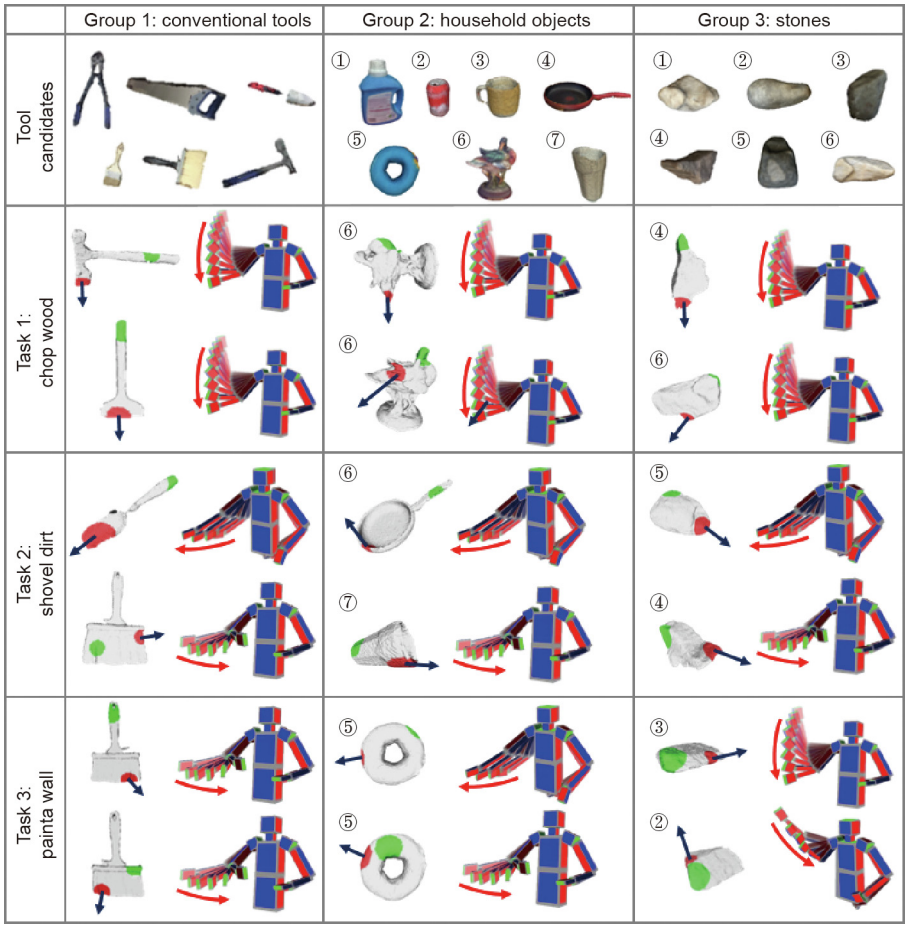

Fig. 22. Given the three tasks of chopping wood, shoveling dirt, and painting a wall, an algorithm proposed by Zhu et al. [232] picks and ranks objects within groups in terms of which object in each group is the best fit for task performance: conventional tools, household objects, and stones. Second, 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.

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 [70,165,181–184], 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.

"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

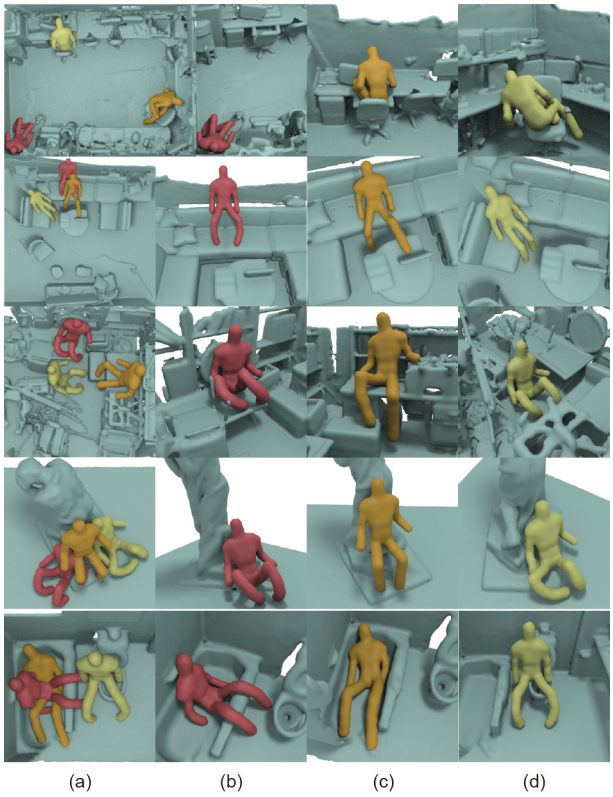

Fig. 23. (a) Top three poses in various scenes for affordance (sitting) recognition. The zoom-in shows views of the (b) best, (c) second-best, and (d) third-best choice of sitting poses. The top two rows are canonical scenarios, the middle row is a cluttered scenario, and the bottom two rows are novel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

with functional grouping/support relationships to build natural and fitting activity spaces.

5.3. Mirroring: Causal-equivalent functionality and 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].

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.

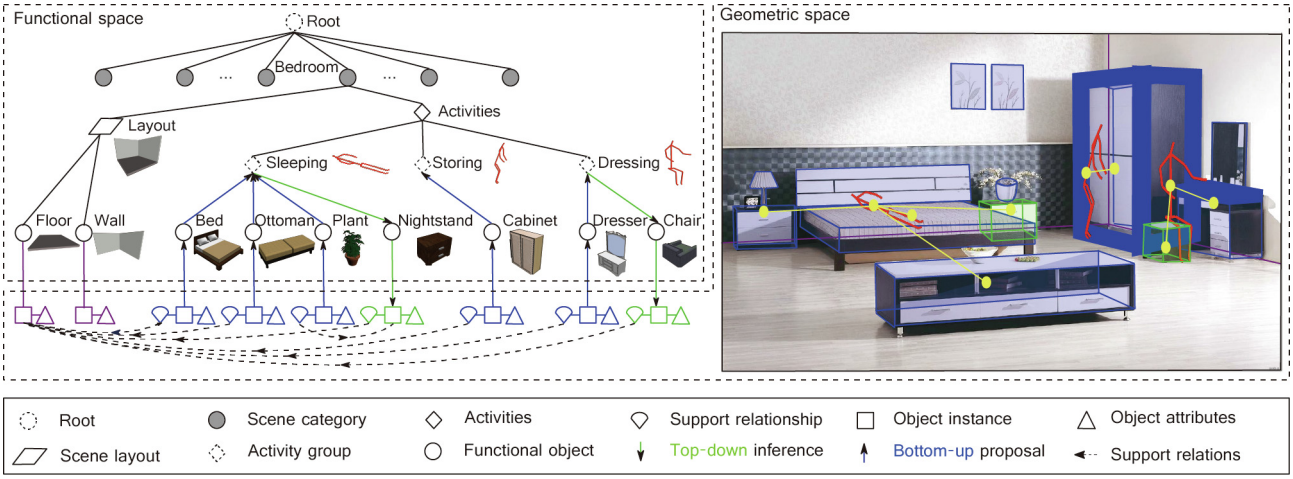

Fig. 24. Task-centered representation of an indoor scene. The functional space exhibits a hierarchical structure, and the geometric space encodes the spatial entities with contextual relationships. The objects are grouped by their hidden activity, i.e., by latent human context or action. Reproduced from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

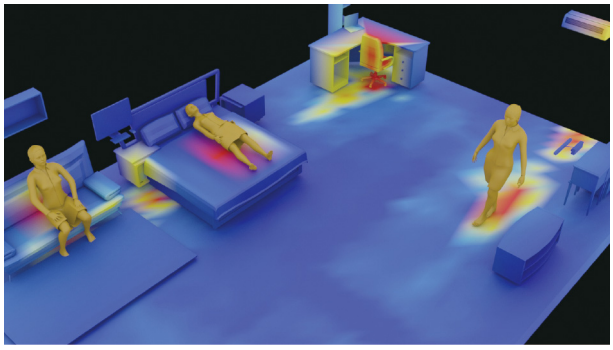

Fig. 25. An example of a synthesized human-centric indoor scene (a bedroom) with an affordance heat map generated by Refs. [99,288]. The joint sampling of the scene was achieved by alternatively sampling humans and objects according to a joint probability distribution.

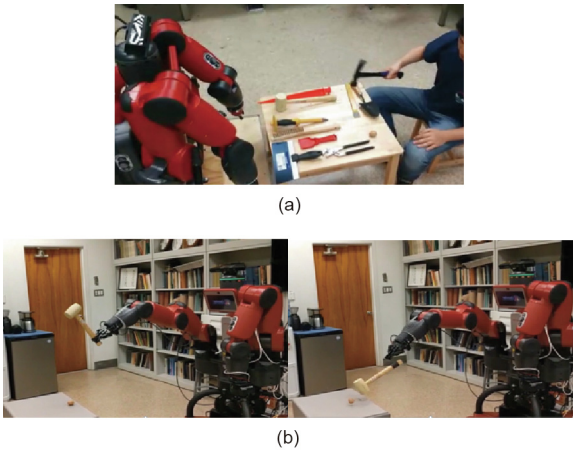

Fig. 26. (a) Given a successful human demonstration, (b) the robot may fail to accomplish the same task by imitating the human demonstration due to different embodiments. In this case, a two-finger gripper cannot firmly hold a hammer while swinging; the hammer slips, and the execution fails.

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 mechanism become observable. From these observations, Edmonds et al. [295,297] 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 [296].

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 over-imitation*: 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.

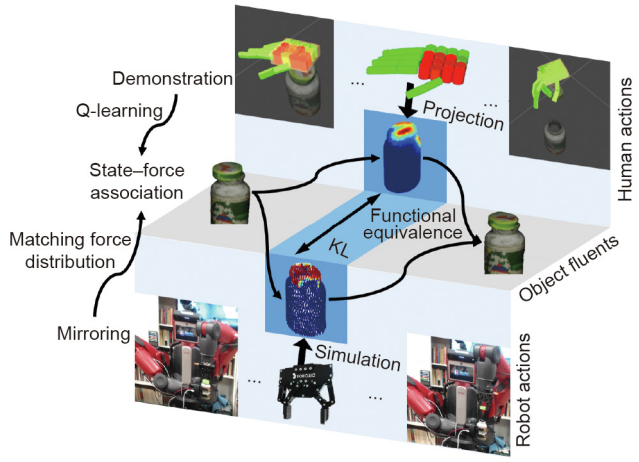

Fig. 27. A robot mirrors human demonstrations with functional equivalence by inferring the action that produces similar force, resulting in similar changes in physical states. *Q*-learning is applied to similar types of forces with categories of object state changes to produce human–object-interaction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

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 ① 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 ② 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].

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 [307,311,312].

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.

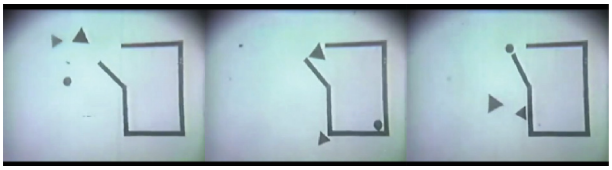

Fig. 28. The seminal Heider–Simmel experiment [313]. Adults can perceive and attribute mental states from nothing but the motion of simple geometric shapes.

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 characters 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: ① They devote their limited time and resources only to those actions that change the world in accordance with their intentions and desires; and ② they achieve their intentions rationally by

maximizing their utility while minimizing their costs, given their beliefs about the world [301,317,318].

Guided by this principle, Gao et al. [319] 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 [319] 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.

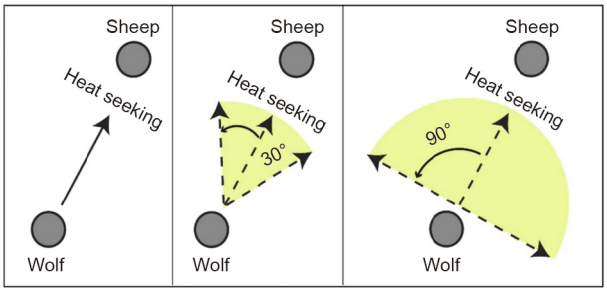

Fig. 29. An illustration of chasing subtlety manipulation in the "Don't Get Caught" experiment. When chasing subtlety is set to zero, the wolf always heads directly toward the (moving) sheep in a "heat-seeking" manner. When the chasing subtlety is set to 30, the wolf always moves in the general direction of the sheep, but is not on a perfect, heat-seeking trajectory; instead, it can move in any direction within a 60° 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. [319] with permission of Elsevier Inc., ©2009.

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 [320,321]. 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. [318] 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 [104,114,322]. 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 [322–325].

Moving from the symbolic input to real video input, Holtzen et al. [326] 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 [326] for more details. The intentions are represented as a

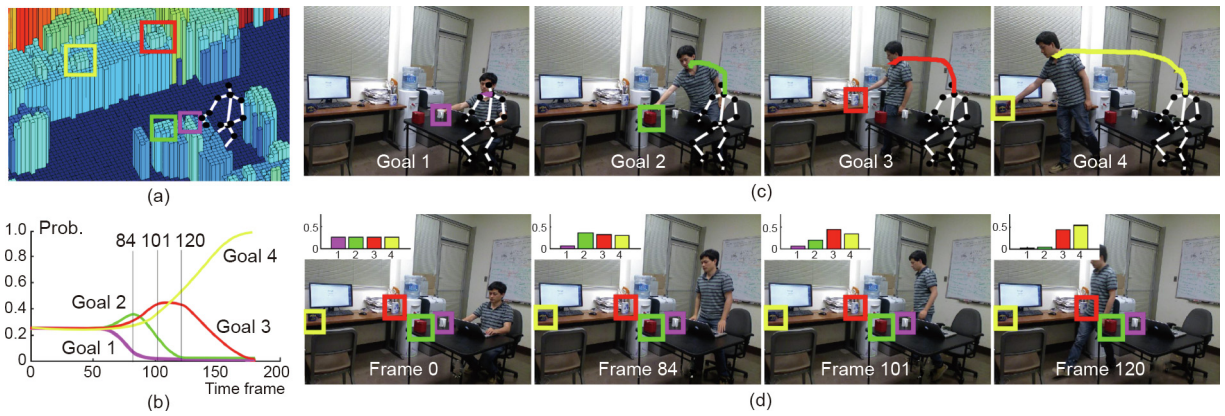

Fig. 30. The plan inference task presented in Ref. [326], seen from the perspective of an observing robot. (a) Four different goals (target objects) in a 3D scene. (b) One outcome of the proposed method: the marginal probability (Prob.) of each terminal action over time. Note that terminal actions are 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. [326] with permission of IEEE, ©2016.

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. 10. 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. [72] 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.

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: ① action-effect association, ② simulation procedures, and ③ teleological reasoning. They concluded that action–effect association and simulation could only serve action monitoring and prediction; social learning, in contrast, requires the inferential productivity of teleological reasoning.

Simulation theory claims that the mechanism underlying the attribution of intentions to actions might rely on simulating the observed action and mapping it onto our own experiences and 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

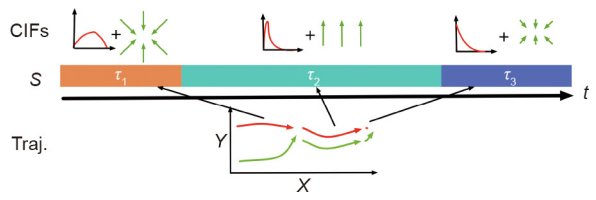

Fig. 31. Inference of human interaction from motion trajectories. The top row demonstrates change within a conditional interactive field (CIF) in sub-interactions as the interaction proceeds, where the CIF models the expected relative motion pattern conditioned on the reference agent's motion. The bottom 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.

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 [102,332]. 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 [334] takes a "functional stance" for the computational representation of goal-directed action [103], where such teleological representations are generated by the aforementioned inferential rationality principle [335]. 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: ① an intuitive sense of whether physical laws are obeyed or violated, and ② 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 [336]. Evidence from psychology suggests that eyes are a cognitively special stimulus with distinctive, "hardwired" pathways in the brain dedicated to their interpretation, revealing

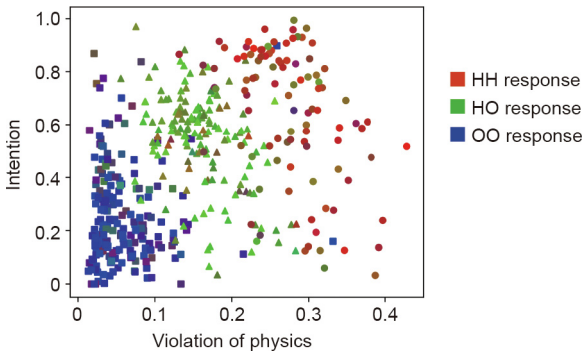

Fig. 32. Constructed psychological space including human–human (HH) animations with 100% animacy degree, human–object (HO) animations, and object–object (OO) animations. Here, a stimulus is depicted by a data point with coordinates derived by the model, and the colors of the data points indicate the 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.

humans' unique ability to infer others' intent from eye gazes [337]. Social eye gaze functions also transcend cultural differences, forming a kind of universal language [338]. 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. [339] 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: ① "Where is the human looking?"—that is, attention/gaze prediction; ② "Why is the human looking?"—that is, intent prediction; and ③ "What task is the human performing?"—that is, task recognition. Wei et al. [339] 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 [339].

Communicative gazes and gestures (e.g., pointing) stand for intent expression and perception in collaborative interactions. Humans need to recognize their partners' communicative intentions in order to collaborate with others and successfully 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 [340] 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. [341] 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 [342] 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

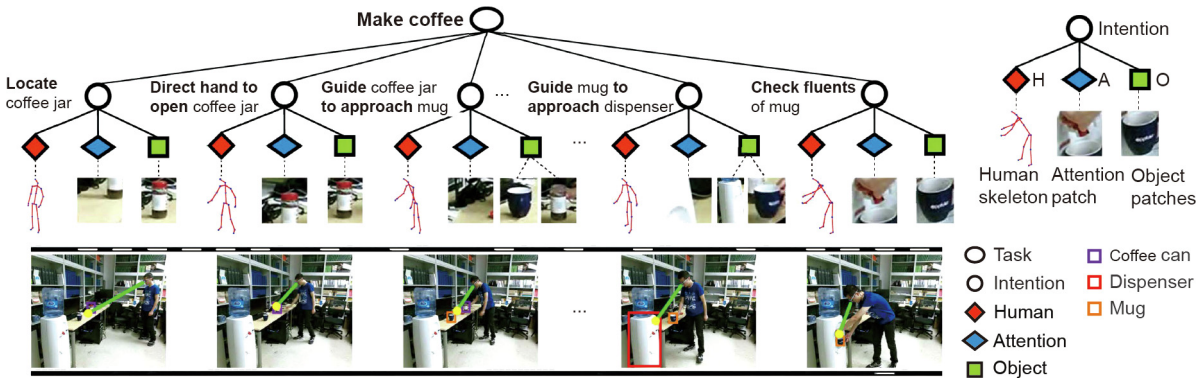

Fig. 33. A task is modeled as sequential intentions in terms of hand–eye coordination with a human–attention–object (HAO) graph. Here, an intention is represented through inverse planning, in which human pose, human attention, and a visible object provide context with which to infer an agent's intention. Reproduced from Ref. [339] with permission of the authors, ©2018.

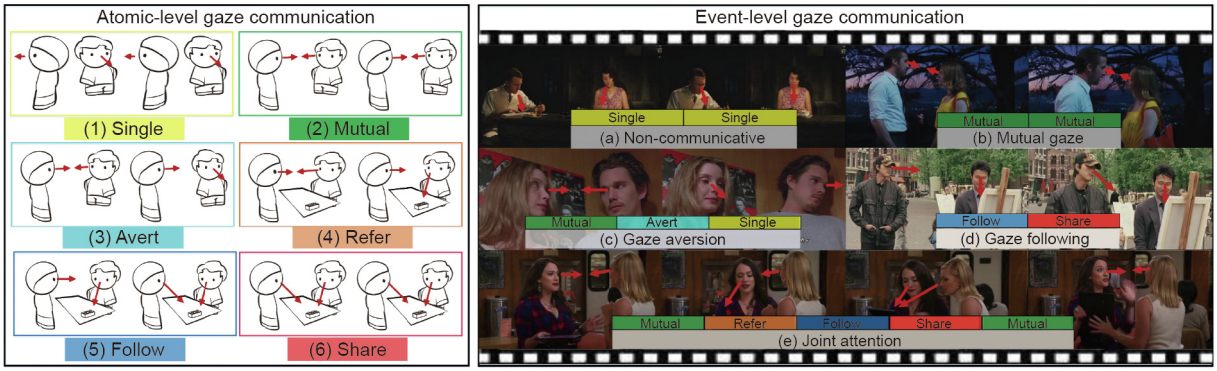

Fig. 34. Human gaze communication dynamics on two hierarchical levels: ① Atomic-level gaze communication describes the fine-grained structures in human gaze interactions; and ② event-level gaze communication refers to long-term social communication events temporally composed of atomic-level gaze communications. Reproduced from Ref. [342] with permission of the authors, ©2019.

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 function 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.

From a formal standpoint, 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

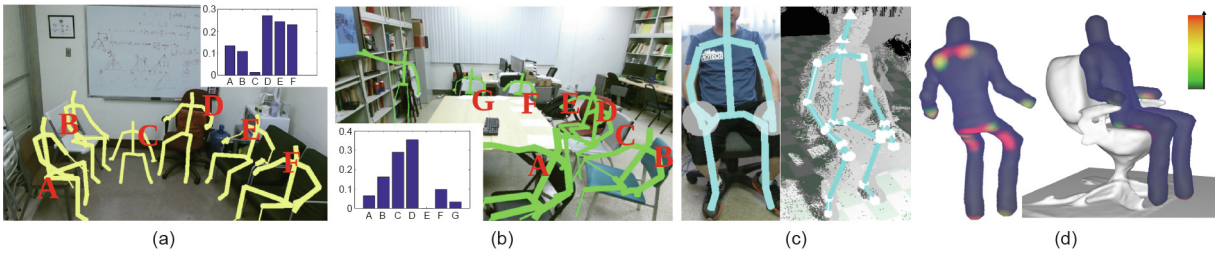

Fig. 35. Examples of sitting in (a) an office and (b) a meeting room. In addition to geometry and appearance, people consider other important factors when deciding where to sit, including comfort level, reaching cost, and social goals. The histograms indicate human preferences for different candidate 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.

accounting for people's preferences in terms of human internal utility.

Similarly, Shukla et al. [349] 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 et al. [350]. The core insight of their work is that language use is a form of rational action; thus, technical tools for reasoning about rational action should elucidate linguistic phenomena [351]. Such a goal-directed view of language production has led to a few interesting language games [352–357], the development of engineering systems for natural language generation [358], and a vocabulary for formal descriptions of pragmatic phenomena in the field of game theory [359,360]. More recently, by assuming the communications between agents to be helpful yet parsimonious, the "Rational Speech Act" [351,361] model has demonstrated promising results in solving some challenging referential games.

By materializing 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. [362] 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) [363,364] and fairness [365] 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 [366].

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:

(1) **A physically realistic VR/mixed reality (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 ① accomplishing the very same task using different sets of objects with different instructions and/or sequences of actions in different environments; and ② 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.

(2) **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 [367], is still a challenging problem. In most cases, algorithms designed for a single agent would be difficult to generalize to a multiple-agent system (MAS) setting [368–370]. We provide a brief review of three related topics in Section 8.2.

(3) **Measuring the limits of an intelligence system: IQ tests.** Studying FPICU opens a new direction of analogy and relational reasoning [371]. Apart from the four-term analogy (or proportional analogy), John C. Raven [372] proposed Raven's progressive matrices (RPM) test in the image domain. The relational and analogical visual reasoning (RAVEN) dataset [373] 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 platforms: 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" [374]. 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 [288,375–377], including MINOS [378], HoME [379], Gibson Environment [380], House3D [381], AI2-THOR [382], VirtualHome [383], VRGym (Fig. 36) [384], and

Fig. 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. [384] with permission of Association for Computing Machinery, © 2019.

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 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 [393–395], 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 FEM 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 [409–411], sand [412,413], rigid bodies [414], fractures [415,416], cloth [417], hair [418], water [419], and solid–fluid

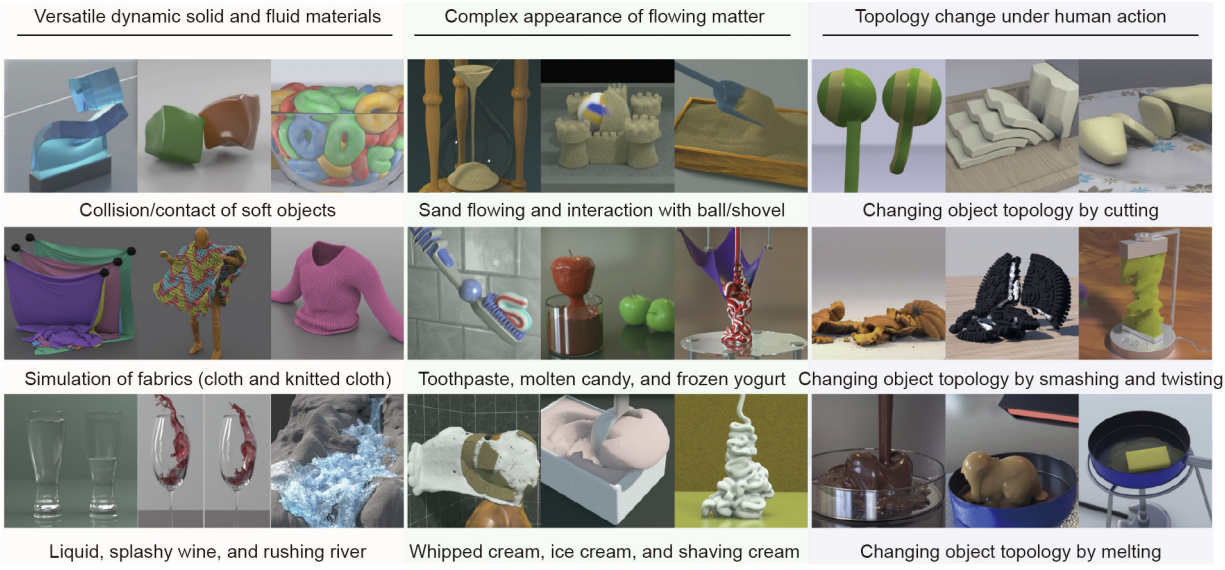

Fig. 37. Diverse physical phenomena simulated using the material point method (MPM).

mixtures [420–422]. 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], hypervelocity 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: The 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 [367]). 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 [370,441] and actor–critic [368,442]-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 [369,443,444] 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 [445–448], 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 [449–451].

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 [366]. 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 [455–457]. 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 [362,366,462].

8.3. Measuring the limits of an 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. [371] 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 [372]—the RPM test.

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 [373]:

(1) 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].

(2) 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.

(3) 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 [373] 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 (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 [373] 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. [373] 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

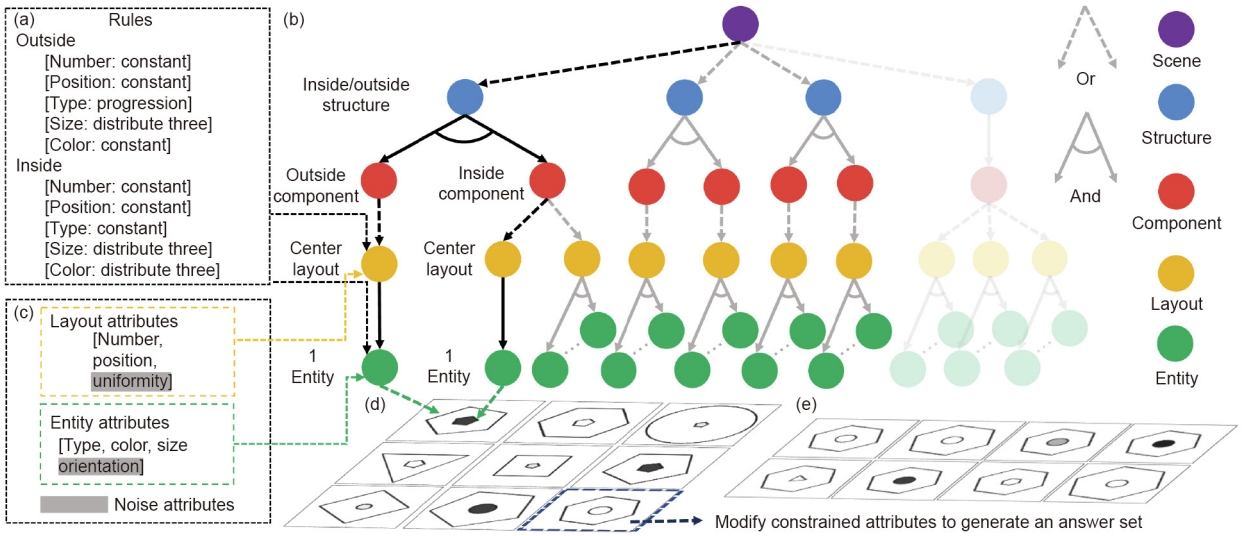

Fig. 38. The RAVEN creation process proposed in Ref. [373]. A graphical illustration of (a) the grammar production rules used in (b) A-SIG. (c) Note that Layout and Entity have associated attributes. (d) A sample problem matrix and (e) a sample candidate set. Reproduced from Ref. [373] with permission of the authors, © 2019.

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 [473–477]. Originating from perceptual learning [478,479], it is well established in the field of psychology and education [480–484] 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.

Acknowledgements

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 United States (in alphabetical order: Carnegie Mellon University, Massachusetts Institute of Technology, Stanford University, University of California at Los Angeles (UCLA), University of Illinois at Urbana–Champaign, and Yale University) and the United Kingdom (in alphabetical order: University of Birmingham, University of Glasgow, University of Leeds, and University of Oxford).[†] The MURI team also holds an annual review meeting at various locations together with two related series of CVPR/CogSci workshops.[‡,††]

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.

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

[‡] Workshop on Vision Meets Cognition: Functionality, Physics, Intentionality, and Causality: https://www.visionmeetscognition.org/.

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

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).

Compliance with ethics guidelines

Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, Joshua B. Tenenbaum, and Song-Chun Zhu declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Marr D. Vision: a computational investigation into the human representation and processing of visual information. San Francisco: W.H. Freeman and Company; 1982.
[2] Mishkin M, Ungerleider LG, Macko KA. Object vision and spatial vision: two cortical pathways. Trends Neurosci 1983;6:414–7.
[3] Ikeuchi K, Hebert M. Task-oriented vision. In: Landy MS, Maloney LT, Pavel M, editors. Exploratory vision. New York: Springer; 1996. p. 257–77.
[4] Land M, Mennie N, Rusted J. The roles of vision and eye movements in the control of activities of daily living. Perception 1999;28(11):1311–28.
[5] Fang F, He S. Cortical responses to invisible objects in the human dorsal and ventral pathways. Nat Neurosci 2005;8(10):1380–5.
[6] Creem-Regehr SH, Lee JN. Neural representations of graspable objects: are tools special? Brain Res Cogn Brain Res 2005;22(3):457–69.
[7] Potter MC. Meaning in visual search. Science 1975;187(4180):965–6.
[8] Potter MC. Short-term conceptual memory for pictures. J Exp Psychol Hum Learn 1976;2(5):509–22.
[9] Schyns PG, Oliva A. From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. Psychol Sci 1994;5(4):195–200.
[10] Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. Nature 1996;381(6582):520–2.
[11] Greene MR, Oliva A. The briefest of glances: the time course of natural scene understanding. Psychol Sci 2009;20(4):464–72.
[12] Greene MR, Oliva A. Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cognit Psychol 2009;58(2):137–76.
[13] Li FF, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene? J Vis 2007;7(1):10.
[14] Rousselet G, Joubert O, Fabre-Thorpe M. How long to get to the "gist" of real-world natural scenes? Vis Cognit 2005;12(6):852–77.
[15] Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 2001;42(3):145–75.
[16] Delorme A, Richard G, Fabre-Thorpe M. Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. Vision Res 2000;40(16):2187–200.
[17] Serre T, Oliva A, Poggio T. A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci USA 2007;104(15):6424–9.
[18] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 2012 Neural Information Processing Systems; 2012 Dec 3–6; Lake Tahoe, NV, USA; 2012.
[19] Kavukcuoglu K, Sermanet P, Boureau YL, Gregor K, Mathieu M, Cun YL. Learning convolutional feature hierarchies for visual recognition. In: Proceedings of the 2010 Neural Information Processing Systems; 2010 Dec 6–11; Vancouver, BC, Canada; 2010.
[20] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA; 2009.
[21] Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. J Neurosci 2018;38(33):7255–69.
[22] Oliva A, Schyns PG. Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Cognit Psychol 1997;34(1):72–107.
[23] Schyns PG. Diagnostic recognition: task constraints, object information, and their interactions. Cognition 1998;67(1–2):147–79.
[24] Malcolm GL, Nuthmann A, Schyns PG. Beyond gist: strategic and incremental information accumulation for scene categorization. Psychol Sci 2014;25(5):1087–97.
[25] Qi S, Huang S, Wei P, Zhu SC. Predicting human activities using stochastic grammar. In: Proceedings of the 2017 IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy; 2017. p. 1164–72.
[26] Pei M, Jia Y, Zhu SC. Parsing video events with goal inference and intent prediction. In: Proceedings of the 2011 IEEE International Conference on Computer Vision; 2011 Nov 6–13; Barcelona, Spain; 2011.
[27] Gosselin F, Schyns PG. Bubbles: a technique to reveal the use of information in recognition tasks. Vision Res 2001;41(17):2261–71.
[28] Ikeuchi K, Hebert M. Task oriented vision. In: Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems; 1992 Jul 7–10; Raleigh, NC, USA; 1992. p. 2187–94.
[29] Hartley R, Zisserman A. Multiple view geometry in computer vision. 2nd ed. Cambridge: Cambridge University Press; 2003.
[30] Ma Y, Soatto S, Kosecka J, Sastry SS. An invitation to 3-D vision: from images to geometric models. New York: Springer Science & Business Media; 2012.
[31] Gupta A, Hebert M, Kanade T, Blei DM. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In: Proceedings of the 2010 Neural Information Processing Systems; 2010 Dec 6–11; Vancouver, BC, Canada; 2010.
[32] Schwing AG, Fidler S, Pollefeys M, Urtasun R. Box in the box: joint 3D layout and object reasoning from single images. In: In: Proceedings of the 2013 IEEE International Conference on Computer Vision; 2013 Dec 1–8; Sydney, Australia. p. 353–60.
[33] Choi W, Chao YW, Pantofaru C, Savarese S. Understanding indoor scenes using 3D geometric phrases. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition; 2013 Jun 25–27; Portland, OR, USA; 2013. p. 33–40.
[34] Zhao Y, Zhu SC. Scene parsing by integrating function, geometry and appearance models. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition; 2013 Jun 25–27; Portland, OR, USA; 2013. p. 3119–26.
[35] Liu X, Zhao Y, Zhu SC. Single-view 3D scene reconstruction and parsing by attribute grammar. IEEE Trans Pattern Anal Mach Intell 2018;40(3):710–25.
[36] Huang S, Qi S, Zhu Y, Xiao Y, Xu Y, Zhu SC. Holistic 3D scene parsing and reconstruction from a single RGB image. In: Proceedings of the 2018 European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany; 2018.
[37] Chen Y, Huang S, Yuan T, Qi S, Zhu Y, Zhu SC. Holistic^{++} scene understanding: single-view 3D holistic scene parsing and human pose estimation with human–object interaction and physical commonsense. In: Proceedings of the 2019 IEEE International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Korea. p. 8648–57.
[38] Huang S, Chen Y, Yuan T, Qi S, Zhu Y, Zhu SC. PerspectiveNet: 3D object detection from a single RGB image via perspective points. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in neural information processing systems 32: proceedings of Neural Information Processing Systems 2019; 2019 Dec 8 14; Vancouver, BC, Canada; 2019. p. 8903 15.
[39] Tolman EC. Cognitive maps in rats and men. Psychol Rev 1948;55(4):189–208.
[40] Wang RF, Spelke ES. Comparative approaches to human navigation. In: Jeffery KJ, editor. The neurobiology of spatial behaviour. Oxford: Oxford University Press; 2003. p. 119–43.
[41] Koenderink JJ, van Doorn AJ, Kappers AM, Lappin JS. Large-scale visual frontoparallels under full-cue conditions. Perception 2002;31(12):1467–75.
[42] Warren WH, Rothman DB, Schnapp BH, Ericson JD. Wormholes in virtual space: from cognitive maps to cognitive graphs. Cognition 2017;166:152–63.
[43] Gillner S, Mallot HA. Navigation and acquisition of spatial knowledge in a virtual maze. J Cogn Neurosci 1998;10(4):445–63.
[44] Foo P, Warren WH, Duchon A, Tarr MJ. Do humans integrate routes into a cognitive map? Map-versus landmark-based navigation of novel shortcuts. J Exp Psychol Learn Mem Cogn 2005;31(2):195–215.
[45] Chrastil ER, Warren WH. From cognitive maps to cognitive graphs. PLoS ONE 2014;9(11):e112544.
[46] Byrne RW. Memory for urban geography. Q J Exp Psychol 1979;31(1):147–54.
[47] Tversky B. Distortions in cognitive maps. Geoforum 1992;23(2):131–8.
[48] Ogle KN. Researches in binocular vision. Philadelphia: WB Saunders; 1950.
[49] Foley JM. Binocular distance perception. Psychol Rev 1980;87(5):411–34.
[50] Luneburg RK. Mathematical analysis of binocular vision. Princeton: Princeton University Press; 1947.
[51] Indow T. A critical review of Luneburg's model with regard to global structure of visual space. Psychol Rev 1991;98(3):430–53.
[52] Gogel WC. A theory of phenomenal geometry and its applications. Percept Psychophys 1990;48(2):105–23.
[53] Glennerster A, Tcheang L, Gilson SJ, Fitzgibbon AW, Parker AJ. Humans ignore motion and stereo cues in favor of a fictional stable world. Curr Biol 2006;16(4):428–32.
[54] Hafting T, Fyhn M, Molden S, Moser MB, Moser EI. Microstructure of a spatial map in the entorhinal cortex. Nature 2005;436(7052):801–6.
[55] Killian NJ, Jutras MJ, Buffalo EA. A map of visual space in the primate entorhinal cortex. Nature 2012;491(7426):761–4.
[56] O'Keefe J, Nadel L. The hippocampus as a cognitive map. Oxford: Clarendon Press; 1978.
[57] Jacobs J, Weidemann CT, Miller JF, Solway A, Burke JF, Wei XX, et al. Direct recordings of grid-like neuronal activity in human spatial navigation. Nat Neurosci 2013;16(9):1188–90.
[58] Fyhn M, Hafting T, Witter MP, Moser EI, Moser MB. Grid cells in mice. Hippocampus 2008;18(12):1230–8.
[59] Doeller CF, Barry C, Burgess N. Evidence for grid cells in a human memory network. Nature 2010;463(7281):657–61.
[60] Yartsev MM, Witter MP, Ulanovsky N. Grid cells without theta oscillations in the entorhinal cortex of bats. Nature 2011;479(7371):103–7.
[61] Gao R, Xie J, Zhu SC, Wu Y. Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion. In: Proceedings of the 2019 International Conference on Learning Representations; 2019 May 6–9; New Orleans, LA, USA; 2019.

[62] Xie J, Gao R, Nijkamp E, Zhu S, Wu YN. Representation learning: a statistical perspective. Annu Rev Stat Appl 2020:7.
[63] Gootjes-Dreesbach L, Pickup LC, Fitzgibbon AW, Glennerster A. Comparison of view-based and reconstruction-based models of human navigational strategy. J Vis 2017;17(9):11.
[64] Vuong J, Fitzgibbon AW, Glennerster A. Human pointing errors suggest a flattened, task-dependent representation of space. bioRxiv 2018:390088.
[65] Choi H, Scholl BJ. Perceiving causality after the fact: postdiction in the temporal dynamics of causal perception. Perception 2006;35(3):385–99.
[66] Scholl BJ, Nakayama K. Illusory causal crescents: misperceived spatial relations due to perceived causality. Perception 2004;33(4):455–69.
[67] Scholl BJ, Gao T. Perceiving animacy and intentionality: visual processing or higher-level judgment. In: Rutherford MD, Kuhlmeier VA, editors. Social perception: detection and interpretation of animacy, agency, and intention. Cambridge: The MIT Press; 2013. p. 197–229.
[68] Scholl BJ. Objects and attention: the state of the art. Cognition 2001;80(1–2):1–46.
[69] Vul E, Alvarez G, Tenenbaum JB, Black MJ. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In: Proceedings of the 2009 Neural Information Processing Systems; 2009 Dec 7–10; Vancouver, BC, Canada; 2009.
[70] Battaglia PW, Hamrick JB, Tenenbaum JB. Simulation as an engine of physical scene understanding. Proc Natl Acad Sci USA 2013;110(45):18327–32.
[71] Hamrick J, Battaglia P, Tenenbaum JB. Internal physics models guide probabilistic judgments about object dynamics. In: Proceedings of the 2011 Annual Meeting of the Cognitive Science Society; 2011 Jul 20–23; Boston, MA, USA; 2011.
[72] Xie D, Shu T, Todorovic S, Zhu SC. Learning and inferring "dark matter" and predicting human intents and trajectories in videos. IEEE Trans Pattern Anal Mach Intell 2018;40(7):1639–52.
[73] Ullman T, Stuhlmüller A, Goodman N, Tenenbaum JB. Learning physics from dynamical scenes. In: Proceedings of the 2014 Annual Meeting of the Cognitive Science Society; 2014 Jul 23–26; Quebec City, QC, Canada; 2014.
[74] Gerstenberg T, Tenenbaum JB. Intuitive theories. In: Waldmann MR, editor. Oxford handbook of causal reasoning. New York: Oxford University Press; 2017. p. 515–48.
[75] Newton I, Colson J. The method of fluxions and infinite series; with its application to the geometry of curve-lines. London: Henry Woodfall; 1736.
[76] Maclaurin C. A treatise of fluxions: in two books. München: Ruddimans; 1742.
[77] Mueller ET. Commonsense reasoning: an event calculus based approach. 2nd ed. Amsterdam: Morgan Kaufmann; 2014.
[78] Mueller ET. Daydreaming in humans and machines: a computer model of the stream of thought. Norwood: Ablex Publishing Corporation; 1990.
[79] Michotte A. The perception of causality. 2nd ed. London: Methuen & Co; 1963.
[80] Carey S. The origin of concepts. New York: Oxford University Press; 2009.
[81] Farhadi A, Endres I, Hoiem D, Forsyth D. Describing objects by their attributes. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA; 2009. p. 1778–85.
[82] Parikh D, Grauman K. Relative attributes. In: Proceedings of the 2011 International Conference on Computer Vision; 2011 Nov 6–13; Barcelona, Spain; 2011. p. 503–10.
[83] Laptev I, Marszałek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. In: Proceedings of the 2008 Conference on Computer Vision and Pattern Recognition; 2008 Jun 24–26; Anchorage, AK, USA; 2008.
[84] Yao B, Zhu SC. Learning deformable action templates from cluttered videos. Proceedings of the 2009 International Conference on Computer Vision; 2009 Sep 29–Oct 2; Kyoto, Japan, 2009.
[85] Yao BZ, Nie BX, Liu Z, Zhu SC. Animated pose templates for modeling and detecting human actions. IEEE Trans Pattern Anal Mach Intell 2013;36(3):436–52.
[86] Wang J, Liu Z, Wu Y, Yuan J. Mining actionlet ensemble for action recognition with depth cameras. Proceedings of the 2012 Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI, USA, 2012.
[87] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 Conference on Computer Vision and Pattern Recognition; 2005 Jun 20–26; San Diego, CA, USA; 2005.
[88] Sadanand S, Corso JJ. Action bank: a high-level representation of activity in video. Proceedings of the 2012 Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI, USA, 2012.
[89] Fleming RW, Barnett-Cowan M, Bülthoff HH. Perceived object stability is affected by the internal representation of gravity. Perception 2010;39:109.
[90] Zago M, Lacquaniti F. Visual perception and interception of falling objects: a review of evidence for an internal model of gravity. J Neural Eng 2005;2(3):S198–208.
[91] Kellman PJ, Spelke ES. Perception of partly occluded objects in infancy. Cognit Psychol 1983;15(4):483–524.
[92] Baillargeon R, Spelke ES, Wasserman S. Object permanence in five-month-old infants. Cognition 1985;20(3):191–208.
[93] Johnson SP, Aslin RN. Perception of object unity in 2-month-old infants. Dev Psychol 1995;31(5):739–45.
[94] Needham A. Factors affecting infants' use of featural information in object segregation. Curr Dir Psychol Sci 1997;6(2):26–33.
[95] Baillargeon R. Infants' physical world. Curr Dir Psychol Sci 2004;13(3):89–94.
[96] Zheng B, Zhao Y, Yu JC, Ikeuchi K, Zhu SC. Detecting potential falling objects by inferring human action and natural disturbance. In: Proceedings of the 2014 International Conference on Robotics and Automation; 2014 May 31–Jun 7; Hong Kong, China; 2014.
[97] Zheng B, Zhao Y, Yu JC, Ikeuchi K, Zhu SC. Beyond point clouds: scene understanding by reasoning geometry and physics. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition; 2013 Jun 23–28; Portland, OR, USA; 2013. p. 3127–34.
[98] Zheng B, Zhao Y, Yu JC, Ikeuchi K, Zhu SC. Scene understanding by reasoning stability and safety. Int J Comput Vis 2015;112(2):221–38.
[99] Qi S, Zhu Y, Huang S, Jiang C, Zhu SC. Human-centric indoor scene synthesis using stochastic grammar. In: Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–22; Salt Lake City, UT, USA; 2018.
[100] Huang S, Qi S, Xiao Y, Zhu Y, Wu YN, Zhu SC. Cooperative holistic scene understanding: unifying 3D object, layout, and camera pose estimation. In: Proceedings of the 2018 Neural Information Processing Systems; 2018 Dec 3–8; Montreal, QC, Canada; 2018.
[101] Gupta A, Satkin S, Efros AA, Hebert M. From 3D scene geometry to human workspace. In: Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition; 2011 Jun 20–25; Providence, RI, USA; 2011.
[102] Iacoboni M, Molnar-Szakacs I, Gallese V, Buccino G, Mazziotta JC, Rizzolatti G. Grasping the intentions of others with one's own mirror neuron system. PLoS Biol 2005;3(3):e79.
[103] Csibra G, Gergely G. 'Obsessed with goals': functions and mechanisms of teleological interpretation of actions in humans. Acta Psychol 2007;124(1):60–78.
[104] Baker CL, Tenenbaum JB, Saxe RR. Goal inference as inverse planning. In: Proceedings of the 2007 Annual Meeting of the Cognitive Science Society; 2007 Aug 1–4; Austin, TX, USA; 2007.
[105] Baker CL, Goodman ND, Tenenbaum JB. Theory-based social goal inference. In: Proceedings of the 2008 Annual Meeting of the Cognitive Science Society; 2008 Jul 23–27; Washington, DC, USA; 2008. p. 1447–52.
[106] Hoai M, De la Torre F. Max-margin early event detectors. Int J Comput Vis 2014;107(2):191–202.
[107] Turek MW, Hoogs A, Collins R. Unsupervised learning of functional categories in video scenes. In: Proceedings of the 2010 European Conference on Computer Vision; 2010 Sep 5–11; Heraklion, Greece. p. 664–77.
[108] Grabner H, Gall J, van Gool L. What makes a chair a chair? In: Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition; 2011 Jun 20–25; Providence, RI, USA; 2011. p. 1529–36.
[109] Jia Z, Gallagher A, Saxena A, Chen T. 3D-based reasoning with blocks, support, and stability. In: Proceedings of the 2013 Conference on Computer Vision and Pattern Recognition; 2013 Jun 23–28; Portland, OR, USA; 2013. p. 1–8.
[110] Jiang Y, Koppula H, Saxena A. Hallucinated humans as the hidden context for labeling 3D scenes. In: Proceedings of the 2013 Conference on Computer Vision and Pattern Recognition; 2013 Jun 23–28; Portland, OR, USA; 2013. p. 2993–3000.
[111] Shu T, Thurman SM, Chen D, Zhu SC, Lu H. Critical features of joint actions that signal human interaction. In: Proceedings of the 2016 Annual Meeting of the Cognitive Science Society; 2016 Aug 10–13; Philadelphia, PA, USA; 2016.
[112] Shu T, Peng Y, Fan L, Lu H, Zhu SC. Perception of human interaction based on motion trajectories: from aerial videos to decontextualized animations. Top Cogn Sci 2018;10(1):225–41.
[113] Shu T, Peng Y, Lu H, Zhu SC. Partitioning the perception of physical and social events within a unified psychological space. In: Proceedings of the 2019 Annual Meeting of the Cognitive Science Society; 2019 Jul 24–27; Montreal, QC, Canada; 2019.
[114] Baker C, Saxe R, Tenenbaum J. Bayesian theory of mind: modeling joint belief-desire attribution. In: Proceedings of the 2011 Annual Meeting of the Cognitive Science Society; 2011 Jul 20–23; Boston, MA, USA; 2011.
[115] Zhao Y, Holtzen S, Gao T, Zhu SC. Represent and infer human theory of mind for human–robot interaction. Proceedings of the 2015 AAAI Fall Symposium Series; 2015 Nov 12–14; Arlington, VA, USA, 2015.
[116] Nisan N, Ronen A. Algorithmic mechanism design. Games Econ Behav 2001;35(1–2):166–96.
[117] Bentham J. An introduction to the principles of morals. London: Athlone; 1935.
[118] Nishant S. Utility learning, non-Markovian planning, and task-oriented programming language [dissertation]. Los Angeles: University of California; 2019.
[119] Robb AA. Optical geometry of motion: a new view of the theory of relativity. W Heffer 1911.
[120] Malament DB. The class of continuous timelike curves determines the topology of spacetime. J Math Phys 1977;18(7):1399–404.
[121] Robb AA. Geometry of time and space. New York: Cambridge University Press; 2014.
[122] Corrigan R, Denton P. Causal understanding as a developmental primitive. Dev Rev 1996;16(2):162–202.
[123] White PA. Causal processing: origins and development. Psychol Bull 1988;104(1):36–52.
[124] Chen YC, Scholl BJ. The perception of history: seeing causal history in static shapes induces illusory motion perception. Psychol Sci 2016;27(6):923–30.
[125] Holyoak KJ, Cheng PW. Causal learning and inference as a rational process: the new synthesis. Annu Rev Psychol 2011;62(1):135–63.

[126] Shanks DR, Dickinson A. Associative accounts of causality judgment. Psychol Learn Motiv 1988;21:229–61.
[127] Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, editors. Classical conditioning II: current research and theory. New York: Appleton-Century-Crofts; 1972. p. 64–99.
[128] Lu H, Yuille AL, Liljeholm M, Cheng PW, Holyoak KJ. Bayesian generic priors for causal learning. Psychol Rev 2008;115(4):955–84.
[129] Edmonds M, Qi S, Zhu Y, Kubricht J, Zhu SC, Lu H. Decomposing human causal learning: bottom-up associative learning and top-down schema reasoning. In: Proceedings of the 2019 Annual Meeting of the Cognitive Science Society; 2019 Jul 24–27; Montreal, QC, Canada; 2019.
[130] Waldmann MR, Holyoak KJ. Predictive and diagnostic learning within causal models: asymmetries in cue competition. J Exp Psychol Gen 1992;121(2):222–36.
[131] Edmonds M, Kubricht J, Summers C, Zhu Y, Rothrock B, Zhu SC, et al. Human causal transfer: challenges for deep reinforcement learning. In: Proceedings of the 2018 Annual Meeting of the Cognitive Science Society; 2018 Jul 25–28; Madison, CT, USA; 2018.
[132] Cheng PW. From covariation to causation: a causal power theory. Psychol Rev 1997;104(2):367–405.
[133] Scholl BJ, Tremoulet PD. Perceptual causality and animacy. Trends Cogn Sci 2000;4(8):299–309.
[134] Rolfs M, Dambacher M, Cavanagh P. Visual adaptation of the perception of causality. Curr Biol 2013;23(3):250–4.
[135] McCollough C. Color adaptation of edge-detectors in the human visual system. Science 1965;149(3688):1115–6.
[136] Kominsky JF, Scholl BJ. Retinotopically specific visual adaptation reveals the structure of causal events in perception. In: Proceedings of the 2018 Annual Meeting of the Cognitive Science Society; 2018 Jul 25–28; Madison, CT, USA; 2018.
[137] Gerstenberg T, Peterson MF, Goodman ND, Lagnado DA, Tenenbaum JB. Eye-tracking causality. Psychol Sci 2017;28(12):1731–44.
[138] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[139] Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region policy optimization. In: Proceedings of the 2015 International Conference on Machine Learning; 2015 Jul 6–11; Lille, France; 2015.
[140] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529(7587):484–9.
[141] Levine S, Finn C, Darrell T, Abbeel P. End-to-end training of deep visuomotor policies. J Mach Learn Res 2016;17(1):1334–73.
[142] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017. arXiv:1707.06347.
[143] Zhang C, Vinyals O, Munos R, Bengio S. A study on overfitting in deep reinforcement learning. 2018. arXiv:1804.06893.
[144] Kansky K, Silver T, Mély DA, Eldawy M, Lázaro-Gredilla M, Lou X, et al. Schema networks: zero-shot transfer with a generative causal model of intuitive physics. 2017. arXiv:1706.04317.
[145] Edmonds M, Ma X, Qi S, Zhu Y, Lu H, Zhu SC. Theory-based causal transfer: integrating instance-level induction and abstract-level structure learning. 2019. arXiv:1911.11185.
[146] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66(5):688–701.
[147] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences. New York: Cambridge University Press; 2015.
[148] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41–55.
[149] Pearl J. Causality: models, reasoning and inference. New York: Cambridge University Press; 2000.
[150] Spirtes P, Glymour C, Scheines R, Heckerman D, Meek C, Cooper GF, et al. Causation, prediction, and search. 2nd ed. Cambridge: MIT Press; 2000.
[151] Chickering DW. Optimal structure identification with greedy search. J Mach Learn Res 2002;3:507–54.
[152] Peters J, Mooij JM, Janzing D, Schölkopf B. Causal discovery with continuous additive noise models. J Mach Learn Res 2014;15(1):2009–53.
[153] He YB, Geng Z. Active learning of causal networks with intervention experiments and optimal designs. J Mach Learn Res 2008;9(11):2523–47.
[154] Bramley NR, Dayan P, Griffiths TL, Lagnado DA. Formalizing Neurath's ship: approximate algorithms for online causal learning. Psychol Rev 2017;124(3):301–38.
[155] Fisher RA. The design of experiments. London: Oliver and Boyd; 1935.
[156] Fire A, Zhu SC. Learning perceptual causality from video. ACM Trans Intell Syst Technol 2016;7(2):23.
[157] Fire A, Zhu SC. Using causal induction in humans to learn and infer causality from video. In: Proceedings of the 2013 Annual Meeting of the Cognitive Science Society; 2013 Jul 31–Aug 3; Berlin, Germany; 2013.
[158] Zhu SC, Wu YN, Mumford D. Minimax entropy principle and its application to texture modeling. Neural Comput 1997;9(8):1627–60.
[159] Xu Y, Qin L, Liu X, Xie J, Zhu SC. A causal and–or graph model for visibility fluent reasoning in tracking interacting objects. In: Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–22; Salt Lake City, UT, USA; 2018. p. 2178–87.
[160] Xiong C, Shukla N, Xiong W, Zhu SC. Robot learning with a spatial, temporal, and causal and–or graph. In: Proceedings of the 2016 IEEE International Conference on Robotics and Automation; 2016 May 16–21; Stockholm, Sweden; 2016.
[161] McCloskey M, Washburn A, Felch L. Intuitive physics: the straight-down belief and its origin. J Exp Psychol Learn Mem Cogn 1983;9(4):636–49.
[162] McCloskey M, Caramazza A, Green B. Curvilinear motion in the absence of external forces: naive beliefs about the motion of objects. Science 1980;210(4474):1139–41.
[163] DiSessa AA. Unlearning Aristotelian physics: a study of knowledge-based learning. Cogn Sci 1982;6(1):37–75.
[164] Kaiser MK, Jonides J, Alexander J. Intuitive reasoning about abstract and familiar physics problems. Mem Cognit 1986;14(4):308–12.
[165] Smith KA, Battaglia P, Vul E. Consistent physics underlying ballistic motion prediction. In: Proceedings of the 2013 Annual Meeting of the Cognitive Science Society; 2013 Jul 31–Aug 3; Berlin, Germany; 2013.
[166] Kaiser MK, Proffitt DR, Whelan SM, Hecht H. Influence of animation on dynamical judgments. J Exp Psychol Hum Percept Perform 1992;18(3):669–89.
[167] Kaiser MK, Proffitt DR, Anderson K. Judgments of natural and anomalous trajectories in the presence and absence of motion. J Exp Psychol Learn Mem Cogn 1985;11(4):795–803.
[168] Kim IK, Spelke ES. Perception and understanding of effects of gravity and inertia on object motion. Dev Sci 1999;2(3):339–62.
[169] Piaget J, Cook MT. The origins of intelligence in children. New York: International Universities Press; 1952.
[170] Piaget J, Cook MT. The construction of reality in the child. New York: Basic Books; 1954.
[171] Hespos SJ, Baillargeon R. Décalage in infants' knowledge about occlusion and containment events: converging evidence from action tasks. Cognition 2006;99(2):B31–41.
[172] Hespos SJ, Baillargeon R. Young infants' actions reveal their developing knowledge of support variables: converging evidence for violation-of-expectation findings. Cognition 2008;107(1):304–16.
[173] Bower TGR. Development in infancy. New York: WH Freeman; 1974.
[174] Leslie AM, Keeble S. Do six-month-old infants perceive causality? Cognition 1987;25(3):265–88.
[175] Luo Y, Baillargeon R, Brueckner L, Munakata Y. Reasoning about a hidden object after a delay: evidence for robust representations in 5-month-old infants. Cognition 2003;88(3):B23–32.
[176] Baillargeon R, Li J, Ng W, Yuan S. An account of infants' physical reasoning. In: Woodward A, Needham A, editors. Learning and the infant mind. New York: Oxford University Press; 2009. p. 66–116.
[177] Baillargeon R. The acquisition of physical knowledge in infancy: a summary in eight lessons. Blackwell Handb Child Cognit Dev 2002;1:46–83.
[178] Achinstein P. The nature of explanation. New York: Oxford University Press; 1983.
[179] Fischer J, Mikhael JG, Tenenbaum JB, Kanwisher N. Functional neuroanatomy of intuitive physical inference. Proc Natl Acad Sci USA 2016;113(34):E5072–81.
[180] Ullman TD, Spelke E, Battaglia P, Tenenbaum JB. Mind games: game engines as an architecture for intuitive physics. Trends Cogn Sci 2017;21(9):649–65.
[181] Bates C, Yildirim I, Tenenbaum JB, Battaglia PW. Humans predict liquid dynamics using probabilistic simulation. In: Proceedings of the 2015 Annual Meeting of the Cognitive Science Society; 2015 Jul 23–25; Pasadena, CA, USA; 2015.
[182] Kubricht J, Jiang C, Zhu Y, Zhu SC, Terzopoulos D, Lu H. Probabilistic simulation predicts human performance on viscous fluid-pouring problem. In: Proceedings of the 2016 Annual Meeting of the Cognitive Science Society; 2016 Aug 10–13; Philadelphia, PA, USA; 2016.
[183] Kubricht J, Zhu Y, Jiang C, Terzopoulos D, Zhu SC, Lu H. Consistent probabilistic simulation underlying human judgment in substance dynamics. In: Proceedings of the 2017 Annual Meeting of the Cognitive Science Society; 2017 Jul 26–29; London, UK; 2017.
[184] Kubricht JR, Holyoak KJ, Lu H. Intuitive physics: current research and controversies. Trends Cogn Sci 2017;21(10):749–59.
[185] Mumford D, Desolneux A. Pattern theory: the stochastic analysis of real-world signals. Boca Raton: CRC Press; 2010.
[186] Mumford D. Pattern theory: a unifying perspective. In: Joseph A, Mignot F, Murat F, Prum B, Rentschler R, editors. First European congress of mathematics. Heidelberg: Springer; 1994. p. 187–224.
[187] Julesz B. Visual pattern discrimination. IRE Trans Inf Theory 1962;8(2):84–92.
[188] Zhu SC, Wu Y, Mumford D. Filters, random fields and maximum entropy (frame): towards a unified theory for texture modeling. Int J Comput Vis 1998;27(2):107–26.
[189] Julesz B. Textons, the elements of texture perception, and their interactions. Nature 1981;290(5802):91–7.
[190] Zhu SC, Guo CE, Wang Y, Xu Z. What are textons? Int J Comput Vis 2005;62(1–2):121–43.
[191] Guo C, Zhu SC, Wu YN. Towards a mathematical theory of primal sketch and sketchability. In: Proceedings of the 9th IEEE International Conference on Computer Vision; 2003 Oct 13–16; Nice, France; 2003.
[192] Guo C, Zhu SC, Wu YN. Primal sketch: integrating structure and texture. Comput Vis Image Underst 2007;106(1):5–19.

[193] Nitzberg M, Mumford DB. The 2.1-D sketch. In: Proceedings of the 3rd International Conference on Computer Vision; 1990 Dec 4–7; Osaka, Japan; 1990.
[194] Wang JYA, Adelson EH. Layered representation for motion analysis. In: Proceedings of the 1993 IEEE Conference on Computer Vision and Pattern Recognition; 1993 Jun 15–17; New York, NY, USA; 1993.
[195] Wang JA, Adelson EH. Representing moving images with layers. IEEE Trans Image Process 1994;3(5):625–38.
[196] Marr D, Nishihara HK. Representation and recognition of the spatial organization of three-dimensional shapes. Proc R Soc Lond B Biol Sci 1978;200(1140):269–94.
[197] Binford I. Visual perception by computer. In: Proceedings of the 1971 IEEE Conference of Systems and Control; 1971 Dec 15–17; Miami Beach, FL, USA; 1971.
[198] Brooks RA. Symbolic reasoning among 3-D models and 2-D images. Artif Intell 1981;17(1–3):285–348.
[199] Kanade T. Recovery of the three-dimensional shape of an object from a single view. Artif Intell 1981;17(1–3):409–60.
[200] Broadbent D. A question of levels: comment on McClelland and Rumelhart. J Exp Psychol Gen 1985;114(2):189–92.
[201] Lowe D. Perceptual organization and visual recognition. Springer Science & Business Media; 1985. Boston.
[202] Pentland AP. Perceptual organization and the representation of natural form. In: Fischler MA, Firschein O, editors. Readings in computer vision. Amsterdam: Elsevier; 1987. p. 680–99.
[203] Wertheimer M. [Experimental studies on the seeing of motion]. Z Psychol Z Angew Psychol 1912;61(3):161–265. German.
[204] Wagemans J, Elder JH, Kubovy M, Palmer SE, Peterson MA, Singh M, et al. A century of Gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. Psychol Bull 2012;138(6):1172–217.
[205] Wagemans J, Feldman J, Gepshtein S, Kimchi R, Pomerantz JR, van der Helm PA, et al. A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. Psychol Bull 2012;138(6):1218–52.
[206] Köhler W. [The physical Gestalten at rest and in steady state]. Braunschweig: Vieweg und Sohn.; 1920. German.
[207] Köhler W. Physical Gestalten. In: Ellis WD, editor. A source book of Gestalt psychology. London: Routledge & Kegan Paul; 1938. p. 17–54.
[208] Wertheimer M. [Investigations in gestalt theory: II. laws of organization in perceptual forms]. Psychol Forsch 1923;4(1):301–50. German.
[209] Wertheimer M. Laws of organization in perceptual forms. In: Ellis WD, editor. A source book of Gestalt psychology. London: Routledge & Kegan Paul; 1938. p. 71–94.
[210] Koffka K. Principles of Gestalt psychology. London: Routledge; 1935.
[211] Waltz D. Understanding line drawings of scenes with shadows. In: Winston PH, Horn B, editors. The psychology of computer vision. New York: McGraw-Hill Companies; 1975.
[212] Barrow HG, Tenenbaum JM. Interpreting line drawings as three-dimensional surfaces. Artif Intell 1981;17(1–3):75–116.
[213] Lowe DG. Three-dimensional object recognition from single two-dimensional images. Artif Intell 1987;31(3):355–95.
[214] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60(2):91–110.
[215] Solso RL, MacLin MK, MacLin OH. Cognitive psychology. 7th ed. New York: Pearson Education; 2005.
[216] Dayan P, Hinton GE, Neal RM, Zemel RS. The Helmholtz machine. Neural Comput 1995;7(5):889–904.
[217] Roberts LG. Machine perception of three-dimensional solids [dissertation]. Cambridge: Massachusetts Institute of Technology; 1963.
[218] Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: detecting and judging objects undergoing relational violations. Cognit Psychol 1982;14(2):143–77.
[219] Blum M, Griffith A, Neumann B. A stability test for configurations of blocks Technical report. Cambridge: Massachusetts Institute of Technology; 1970.
[220] Brand M, Cooper P, Birnbaum L. Seeing physics, or: physics is for prediction. In: Proceedings of the Workshop on Physics-based Modeling in Computer Vision; 1995 Jun 18–19; Cambridge, MA, USA; 1995. p. 144–50.
[221] Gupta A, Efros AA, Hebert M. Blocks world revisited: image understanding using qualitative geometry and mechanics. In: Proceedings of the 2010 European Conference on Computer Vision; 2010 Sep 5–11; Heraklion, Greece; 2010. p. 482–96.
[222] Hedau V, Hoiem D, Forsyth D. Recovering the spatial layout of cluttered rooms. In: Proceedings of the 2009 International Conference on Computer Vision; 2009 Sep 29–Oct 2; Kyoto, Japan; 2009. p. 1849–56.
[223] Lee DC, Hebert M, Kanade T. Geometric reasoning for single image structure recovery. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA; 2009. p. 2136–43.
[224] Hedau V, Hoiem D, Forsyth D. Recovering free space of indoor scenes from a single image. In: Proceedings of the 2012 Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI, USA; 2012. p. 2807–14.
[225] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGBD images. In: Proceedings of the 2012 European Conference on Computer Vision; 2012 Oct 7–13; Florence, Italy; 2012. p. 746–60.
[226] Schwing AG, Hazan T, Pollefeys M, Urtasun R. Efficient structured prediction for 3D indoor scene understanding. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI, USA; 2012. p. 2815–22.
[227] Guo R, Hoiem D. Support surface prediction in indoor scenes. In: Proceedings of the 2013 IEEE International Conference on Computer Vision; 2013 Dec 1–8; Sydney, NSW, Australia; 2013. p. 2144–51.
[228] Shao T, Monszpart A, Zheng Y, Koo B, Xu W, Zhou K, et al. Imagining the unseen: stability-based cuboid arrangements for scene understanding. ACM Trans Graph 2014;33(6):1–11.
[229] Du Y, Liu Z, Basevi H, Leonardis A, Freeman B, Tenenbaum J, et al. Learning to exploit stability for 3D scene parsing. In: Proceedings of the 2018 Neural Information Processing Systems; 2018 Dec 3–8; Montreal, QC, Canada; 2018.
[230] Wu J, Yildirim I, Lim JJ, Freeman B, Tenenbaum J. Galileo: perceiving physical object properties by integrating a physics engine with deep learning. In: Proceedings of the 2015 Neural Information Processing Systems; 2015 Dec 7–12; Montreal, QC, Canada; 2015.
[231] Wu J, Lim JJ, Zhang H, Tenenbaum JB, Freeman WT. Physics 101: learning physical object properties from unlabeled videos. In: Proceedings of the 2016 British Machine Vision Conference; 2016 Sep 19–22; York, UK; 2016.
[232] Zhu Y, Zhao Y, Zhu SC. Understanding tools: task-oriented object modeling, learning and recognition. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA; 2015. p. 2855–64.
[233] Zhu Y, Jiang C, Zhao Y, Terzopoulos D, Zhu SC. Inferring forces and learning human utilities from videos. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 26–Jul 1; Las Vegas, NV, USA; 2016.
[234] Brubaker MA, Fleet DJ. The kneed walker for human pose tracking. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition; 2008 Jun 23–28; Anchorage, AK, USA; 2008. p. 1–8.
[235] Brubaker MA, Sigal L, Fleet DJ. Estimating contact dynamics. In: Proceedings of the 2009 IEEE International Conference on Computer Vision; 2009 Sep 29–Oct 2; Kyoto, Japan; 2009. p. 2389–96.
[236] Brubaker MA, Fleet DJ, Hertzmann A. Physics-based person tracking using the anthropomorphic walker. Int J Comput Vis 2010;87(1–2):140–55.
[237] Pham TH, Kheddar A, Qammaz A, Argyros AA. Towards force sensing from vision: observing hand-object interactions to infer manipulation forces. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA. p. 2810–9.
[238] Wang Y, Min J, Zhang J, Liu Y, Xu F, Dai Q, et al. Video-based hand manipulation capture through composite motion control. ACM Trans Graph 2013;32(4):43.
[239] Zhao W, Zhang J, Min J, Chai J. Robust realtime physics-based motion control for human grasping. ACM Trans Graph 2013;32(6):207.
[240] Gibson JJ. The perception of the visual world. Boston: Houghton Mifflin; 1950.
[241] Gibson JJ. The senses considered as perceptual systems. Boston: Houghton Mifflin; 1966.
[242] Nelson K. Concept, word, and sentence: interrelations in acquisition and development. Psychol Rev 1974;81(4):267–85.
[243] Gibson JJ. The theory of affordances. In: Gieseking JJ, Mangold W, Katz C, Low S, Saegert S, editors. The people, place, and space reader. New York: Routledge; 2014.
[244] Hassanin M, Khan S, Tahtali M. Visual affordance and function understanding: a survey. 2018. arXiv:1807.06775.
[245] Min H, Yi C, Luo R, Zhu J, Bi S. Affordance research in developmental robotics: a survey. IEEE Trans Cogn Dev Syst 2016;8(4):237–55.
[246] Bohg J, Morales A, Asfour T, Kragic D. Data-driven grasp synthesis—a survey. IEEE Trans Robot 2014;30(2):289–309.
[247] Yamanobe N, Wan W, Ramirez-Alpizar IG, Petit D, Tsuji T, Akizuki S, et al. A brief review of affordance in robotic manipulation research. Adv Robot 2017;31(19–20):1086–101.
[248] Kohler W. The mentality of apes. New York: Routledge; 1925.
[249] Thorpe WH. Learning and instinct in animals. Cambridge: Harvard University Press; 1956.
[250] Oakley KP. Man the tool-maker. Chicago: University of Chicago Press; 1968.
[251] Goodall J. The chimpanzees of Gombe: patterns of behavior. Cambridge: Bellknap Press of the Harvard University Press; 1986.
[252] Whiten A, Goodall J, McGrew WC, Nishida T, Reynolds V, Sugiyama Y, et al. Cultures in chimpanzees. Nature 1999;399(6737):682–5.
[253] Byrne R, Whiten A, editors. Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans. New York: Oxford University Press; 1988.
[254] Santos LR, Rosati A, Sproul C, Spaulding B, Hauser MD. Means-means-end tool choice in cotton-top tamarins (*Saguinus oedipus*): finding the limits on primates' knowledge of tools. Anim Cogn 2005;8:236–46.
[255] Hunt GR. Manufacture and use of hook-tools by New Caledonian crows. Nature 1996;379(6562):249–51.
[256] Weir AA, Chappell J, Kacelnik A. Shaping of hooks in New Caledonian crows. Science 2002;297(5583):981.
[257] McCoy DE, Schiestl M, Neilands P, Hassall R, Gray RD, Taylor AH. New Caledonian crows behave optimistically after using tools. Curr Biol 2019;29(16):2737–42.
[258] Beck BB. Animal tool behavior: the use and manufacture of tools by animals. New York: Garland STPM Press; 1980.

[259] Bird CD, Emery NJ. Insightful problem solving and creative tool modification by captive nontool-using rooks. Proc Natl Acad Sci USA 2009;106(25):10370–5.
[260] Freeman P, Newell A. A model for functional reasoning in design. In: Proceedings of the 1971 International Joint Conference on Artificial Intelligence; 1971 Sep 1–3; London, England; 1971.
[261] Winston PH. Learning structural descriptions from examples Technical report. Cambridge: Massachusetts Institute of Technology; 1970.
[262] Winston PH, Binford TO, Katz B, Lowry M. Learning physical descriptions from functional definitions, examples, and precedents. Proceedings of the 1983 AAAI Conference on Artificial Intelligence; 1983 Aug 22–26; Washington, DC, USA, 1983.
[263] Brady M, Agre PE. The mechanic's mate. In: Proceedings of the 6th European Conference on Artificial Intelligence; 1984 Sep 5–7; Pisa, Italy; 1984. p. 79–94
[264] Connell JH, Brady M. Generating and generalizing models of visual objects. Artif Intell 1987;31(2):159–83.
[265] Ho SB. Representing and using functional definitions for visual recognition [dissertation]. Madison: The University of Wisconsin-Madison; 1987.
[266] DiManzo M, Trucco E, Giunchiglia F, Ricci F. FUR: understanding functional reasoning. Int J Intell Syst 1989;4(4):431–57.
[267] Minsky M. The society of mind. New York: Simon and Schuster Paperbacks; 1988.
[268] Stark L, Bowyer K. Achieving generalized object recognition through reasoning about association of function to structure. IEEE Trans Pattern Anal Mach Intell 1991;13(10):1097–104.
[269] Liu Z, Freeman WT, Tenenbaum JB, Wu J. Physical primitive decomposition. In: Proceedings of the 2018 European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany; 2018.
[270] Baber C. Cognition and tool use: forms of engagement in human and animal use of tools. London: CRC Press; 2003.
[271] Inhelder B, Piaget J. The growth of logical thinking from childhood to adolescence: an essay on the construction of formal operational structures. London: Psychology Press; 1958.
[272] Hespos SJ, Baillargeon R. Reasoning about containment events in very young infants. Cognition 2001;78(3):207–45.
[273] Wang SH, Baillargeon R, Paterson S. Detecting continuity violations in infancy: a new account and new evidence from covering and tube events. Cognition 2005;95(2):129–73.
[274] Hespos SJ, Spelke ES. Precursors to spatial language: the case of containment. In: Aurnague M, Hickmann M, editors. The categorization of spatial entities in language and cognition. Amsterdam: John Benjamins Publishing; 2007. p. 233–45.
[275] Strickland B, Scholl BJ. Visual perception involves event-type representations: the case of containment versus occlusion. J Exp Psychol Gen 2015;144(3):570–80.
[276] Casasola M, Cohen LB. Infant categorization of containment, support and tight-fit spatial relationships. Dev Sci 2002;5(2):247–64.
[277] Davis E, Marcus G, Frazier-Logue N. Commonsense reasoning about containers using radically incomplete information. Artif Intell 2017;248:46–84.
[278] Davis E. How does a box work? A study in the qualitative dynamics of solid objects. Artif Intell 2011;175(1):299–345.
[279] Davis E. Pouring liquids: a study in commonsense physical reasoning. Artif Intell 2008;172(12–13):1540–78.
[280] Cohn AG. Qualitative spatial representation and reasoning techniques. In: Proceedings of the 1997 Annual Conference on Artificial Intelligence; 1997 Sep 9–12; Freiburg, Germany; 1997. p. 1–30.
[281] Cohn AG, Hazarika SM. Qualitative spatial representation and reasoning: an overview. Fundam Inform 2001;46(1–2):1–29.
[282] Liang W, Zhao Y, Zhu Y, Zhu SC. Evaluating human cognition of containing relations with physical simulation. In: Proceedings of the 2015 Annual Meeting of the Cognitive Science Society; 2015 Jul 23–25; Pasadena, CA, USA; 2015.
[283] Yu LF, Duncan N, Yeung SK. Fill and transfer: a simple physics-based approach for containability reasoning. In: Proceedings of the 2015 International Conference on Computer Vision; 2015 Dec 11–18; Santiago, Chile; 2015.
[284] Mottaghi R, Schenck C, Fox D, Farhadi A. See the glass half full: reasoning about liquid containers, their volume and content. In: Proceedings of the 2017 International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy; 2017.
[285] Liang W, Zhao Y, Zhu Y, Zhu SC. What is where: inferring containment relations from videos. In: Proceedings of the 2016 International Joint Conference on Artificial Intelligence; 2016 Jul 9–15; New York, NY, USA; 2016.
[286] Liang W, Zhu Y, Zhu SC. Tracking occluded objects and recovering incomplete trajectories by reasoning about containment relations and human actions. In: Proceedings of the 2018 AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA; 2018.
[287] Jiang Y, Lim M, Saxena A. Learning object arrangements in 3D scenes using human context. In: Proceedings of the 29th International Conference on Machine Learning; 2012 Jun 26–Jul 1; Edinburgh, Scotland. p. 907–14.
[288] Jiang C, Qi S, Zhu Y, Huang S, Lin J, Yu LF, et al. Configurable 3D scene synthesis and 2D image rendering with per-pixel ground truth using stochastic grammars. Int J Comput Vis 2018;126(9):920–41.
[289] Dautenhahn K, Nehaniv CL, editors. Imitation in animals and artifacts. Cambridge: MIT Press; 2002.
[290] Argall BD, Chernova S, Veloso M, Browning B. A survey of robot learning from demonstration. Robot Auton Syst 2009;57(5):469–83.
[291] Osa T, Pajarinen J, Neumann G, Bagnell JA, Abbeel P, Peters J. An algorithmic perspective on imitation learning. Found Trends Rob 2018;7(1–2):1–179.
[292] Gu Y, Sheng W, Liu M, Ou Y. Fine manipulative action recognition through sensor fusion. In: Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2015 Sep 28–Oct 2; Hamburg, Germany; 2015.
[293] Hammond FL, Mengüç Y, Wood RJ. Toward a modular soft sensor-embedded glove for human hand motion and tactile pressure measurement. In: Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2014 Sep 14–18; Chicago, IL, USA. p. 4000–7.
[294] Liu H, Xie X, Millar M, Edmonds M, Gao F, Zhu Y, et al. A glove-based system for studying hand-object manipulation via joint pose and force sensing. In: Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2017 Sep 24–28; Vancouver, BC, USA. p. 6617–24.
[295] Edmonds M, Gao F, Xie X, Liu H, Qi S, Zhu Y, et al. Feeling the force: integrating force and pose for fluent discovery through imitation learning to open medicine bottles. In: Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2017 Sep 24–28; Vancouver, BC, USA. p. 3530–7.
[296] Liu H, Zhang Y, Si W, Xie X, Zhu Y, Zhu SC. Interactive robot knowledge patching using augmented reality. In: Proceedings of the 2018 IEEE International Conference on Robotics and Automation; 2018 May 21–25; Brisbane, QLD, Australia. p. 1947–54.
[297] Edmonds M, Gao F, Liu H, Xie X, Qi S, Rothrock B, et al. A tale of two explanations: enhancing human trust by explaining robot behavior. Sci Robot 2019;4(37):eaay4663.
[298] Liu H, Zhang C, Zhu Y, Jiang C, Zhu SC. Mirroring without overimitation: learning functionally equivalent manipulation actions. Proceedings of the 2019 AAAI Conference on Artificial Intelligence; 2019 Jan 27–Feb 1; Honolulu, HI, USA, 2019.
[299] Dennett DC. The intentional stance. Cambridge: MIT Press; 1989.
[300] Heider F. The psychology of interpersonal relations. London: Psychology Press; 2013.
[301] Gergely G, Nádasdy Z, Csibra G, Bíró S. Taking the intentional stance at 12 months of age. Cognition 1995;56(2):165–93.
[302] Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behav Brain Sci 1978;1(4):515–26.
[303] Baldwin DA, Baird JA. Discerning intentions in dynamic human action. Trends Cogn Sci 2001;5(4):171–8.
[304] Woodward AL. Infants selectively encode the goal object of an actor's reach. Cognition 1998;69(1):1–34.
[305] Meltzoff AN, Brooks R. "Like me" as a building block for understanding other minds: bodily acts, attention, and intention. In: Malle BF, Moses LJ, Baldwin DA, editors. Intentions and intentionality: foundations of social cognition. Cambridge: MIT Press; 2001. p. 171–92.
[306] Baldwin DA, Baird JA, Saylor MM, Clark MA. Infants parse dynamic action. Child Dev 2001;72(3):708–17.
[307] Tomasello M, Carpenter M, Call J, Behne T, Moll H. Understanding and sharing intentions: the origins of cultural cognition. Behav Brain Sci 2005;28(5):675–91.
[308] Biro S, Hommel B. Becoming an intentional agent: introduction to the special issue. Acta Psychol 2007;124(1):1–7.
[309] Gergely G, Bekkering H, Király I. Rational imitation in preverbal infants. Nature 2002;415(6873):755.
[310] Woodward AL, Sommerville JA, Gerson S, Henderson AME, Buresh J. The emergence of intention attribution in infancy. Psychol Learn Motiv 2009;51:187–222.
[311] Zelazo PD, Astington JW, Olson DR, editors. Developing theories of intention: social understanding and self-control. Mahwah: Lawrence Erlbaum Associates Publishers; 1999.
[312] Bloom P. Intention, history, and artifact concepts. Cognition 1996;60(1):1–29.
[313] Heider F, Simmel M. An experimental study of apparent behavior. Am J Psychol 1944;57(2):243–59.
[314] Berry DS, Misovich SJ. Methodological approaches to the study of social event perception. Pers Soc Psychol Bull 1994;20(2):139–52.
[315] Bassili JN. Temporal and spatial contingencies in the perception of social events. J Pers Soc Psychol 1976;33(6):680–5.
[316] Dittrich WH, Lea SE. Visual perception of intentional motion. Perception 1994;23(3):253–68.
[317] Dennett DC. Précis of the intentional stance. Behav Brain Sci 1988;11(3):495–505.
[318] Liu S, Brooks NB, Spelke ES. Origins of the concepts cause, cost, and goal in prereaching infants. Proc Natl Acad Sci USA 2019;116(36):17747–52.
[319] Gao T, Newman GE, Scholl BJ. The psychophysics of chasing: a case study in the perception of animacy. Cognit Psychol 2009;59(2):154–79.
[320] Liu S, Spelke ES. Six-month-old infants expect agents to minimize the cost of their actions. Cognition 2017;160:35–42.
[321] Gergely G, Csibra G. Teleological reasoning in infancy: the naïve theory of rational action. Trends Cogn Sci 2003;7(7):287–92.
[322] Baker CL, Saxe R, Tenenbaum JB. Action understanding as inverse planning. Cognition 2009;113(3):329–49.

[323] Pereira LM, Anh HT. Intention recognition via causal Bayes networks plus plan generation. In: Proceedings of the 14th Portuguese Conference on Artificial Intelligence; 2009 Oct 12–15; Aveiro, Portugal; 2009. p. 138–49.
[324] Narang S, Best A, Manocha D. Inferring user intent using Bayesian theory of mind in shared avatar-agent virtual environments. IEEE Trans Vis Comput Graph 2019;25(5):2113–22.
[325] Nakahashi R, Baker CL, Tenenbaum JB. Modeling human understanding of complex intentional action with a Bayesian nonparametric subgoal model. Proceedings of the 2016 AAAI Conference on Artificial Intelligence; 2016 Feb 12–17; Phoenix, AZ, USA, 2016.
[326] Holtzen S, Zhao Y, Gao T, Tenenbaum JB, Zhu SC. Inferring human intent from video by sampling hierarchical plans. In: Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2016 Oct 9–14; Daejeon, Korea. p. 1489–96.
[327] Kong Y, Fu Y. Human action recognition and prediction: a survey. 2018. arXiv:1806.11230.
[328] Blakemore SJ, Decety J. From the perception of action to the understanding of intention. Nat Rev Neurosci 2001;2(8):561–7.
[329] Elsner B, Hommel B. Effect anticipation and action control. J Exp Psychol Hum Percept Perform 2001;27(1):229–40.
[330] Elsner B. Infants' imitation of goal-directed actions: the role of movements and action effects. Acta Psychol 2007;124(1):44–59.
[331] Rizzolatti G, Craighero L. The mirror–neuron system. Annu Rev Neurosci 2004;27(1):169–92.
[332] Kaplan JT, Iacoboni M. Getting a grip on other minds: mirror neurons, intention understanding, and cognitive empathy. Soc Neurosci 2006;1(3–4):175–83.
[333] Reid VM, Csibra G, Belsky J, Johnson MH. Neural correlates of the perception of goal-directed action in infants. Acta Psychol 2007;124(1):129–38.
[334] Csibra G, Gergely G. The teleological origins of mentalistic action explanations: a developmental hypothesis. Dev Sci 2002;1(2):255–9.
[335] Gergely G. The development of understanding self and agency. In: Goswami U, editor. Blackwell handbook of childhood cognitive development. Oxford: Blackwell Publishers Ltd.; 2002. p. 26–46.
[336] Kleinke CL. Gaze and eye contact: a research review. Psychol Bull 1986;100(1):78–100.
[337] Emery NJ. The eyes have it: the neuroethology, function and evolution of social gaze. Neurosci Biobehav Rev 2000;24(6):581–604.
[338] Burgoon JK, Guerrero LK, Floyd K. Nonverbal communication. New York: Routledge; 2016.
[339] Wei P, Liu Y, Shu T, Zheng N, Zhu SC. Where and why are they looking? Jointly inferring human attention and intentions in complex tasks. In: Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–22; Salt Lake City, UT, USA; 2018. p. 6801–9.
[340] Melis AP, Tomasello M. Chimpanzees (*Pan troglodytes*) coordinate by communicating in a collaborative problem-solving task. Proc R Soc B 1901;2019(286):20190408.
[341] Fan L, Chen Y, Wei P, Wang W, Zhu SC. Inferring shared attention in social scene videos. In: Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–22; Salt Lake City, UT, USA; 2018. p. 6460–8.
[342] Fan L, Wang W, Huang S, Tang X, Zhu SC. Understanding human gaze communication by spatio-temporal graph reasoning. In: Proceedings of the 2019 International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Korea. p. 5724–33.
[343] Trick S, Koert D, Peters J, Rothkopf C. Multimodal uncertainty reduction for intention recognition in human–robot interaction. 2019. arXiv:1907.02426.
[344] Shu T, Ryoo MS, Zhu SC. Learning social affordance for human–robot interaction. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence; 2016 Jul 9–15; New York, NY, USA; 2016. p. 3454–61.
[345] Shu T, Gao X, Ryoo MS, Zhu SC. Learning social affordance grammar from videos: transferring human interactions to human–robot interactions. In: Proceedings of the 2017 IEEE International Conference on Robotics and Automation; 2017 May 29–Jun 3; Singapore, Singapore; 2017.
[346] Russell SJ, Norvig P. Artificial intelligence: a modern approach. 3rd ed. New York: Pearson Education Limited; 2016.
[347] Hutcheson F. An inquiry into the original of our ideas of beauty and virtue: in two treatises. 2nd ed. London: Darby J, Bettesworth A, Fayram F, Pemberton J, Rivington C, Hooke J, Clay F, Batley J, Symon E; 1726.
[348] Mill JS. Utilitarianism. 12th ed. New York: Longmans, Green and Company; 1895.
[349] Shukla N, He Y, Chen F, Zhu SC. Learning human utility from video demonstrations for deductive planning in robotics. In: Proceedings of the 1st Annual Conference on Robot Learning; 2017 Nov 13–15; Mountain View, CA, USA. p. 448–57.
[350] Grice HP, Cole P, Morgan J. Logic and conversation. In: Ezcurdia M, Stainton RJ, editors. The semantics–pragmatics boundary in philosophy. Toronto: Broadview Press; 2013.
[351] Goodman ND, Frank MC. Pragmatic language interpretation as probabilistic inference. Trends Cogn Sci 2016;20(11):818–29.
[352] Lewis D. Convention: a philosophical study. Oxford: Blackwell Publishers; 2002.
[353] Sperber D, Wilson D. Relevance: communication and cognition. Cambridge: Harvard University Press; 1986.
[354] Wittgenstein L. Philosophical investigations. New York: Macmillan; 1953.
[355] Clark HH. Using language. Cambridge: Cambridge University Press; 1996.
[356] Qing C, Franke M. Variations on a Bayesian theme: comparing Bayesian models of referential reasoning. In: Zeevat H, Schmitz HC, editors. Bayesian natural language semantics and pragmatics. Heidelberg: Springer; 2015. p. 201–20.
[357] Goodman ND, Stuhlmüller A. Knowledge and implicature: modeling language understanding as social cognition. Top Cogn Sci 2013;5(1):173–84.
[358] Dale R, Reiter E. Computational interpretations of the Gricean maxims in the generation of referring expressions. Cogn Sci 1995;19(2):233–63.
[359] Benz A, Jäger G, van Rooij R. An introduction to game theory for linguists. In: Benz A, Jäger G, van Rooij R, editors. Game theory and pragmatics. London: Palgrave Macmillan; 2006. p. 1–82.
[360] Jäger G. Applications of game theory in linguistics. Lang Linguist Compass 2008;2(3):406–21.
[361] Frank MC, Goodman ND. Predicting pragmatic reasoning in language games. Science 2012;336(6084):998.
[362] Kleiman-Weiner M, Gerstenberg T, Levine S, Tenenbaum JB. Inference of intention and permissibility in moral decision making. In: Proceedings of the 2015 Annual Meeting of the Cognitive Science Society; 2015 Jul 23–25; Pasadena, CA, USA; 2015.
[363] Kleiman-Weiner M, Ho MK, Austerweil JL, Littman ML, Tenenbaum JB. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In: Proceedings of the 2016 Annual Meeting of the Cognitive Science Society; 2016 Aug 10–13; Philadelphia, PA, USA; 2016.
[364] Shum M, Kleiman-Weiner M, Littman ML, Tenenbaum JB. Theory of minds: understanding behavior in groups through inverse planning. In: Proceedings of the 2019 AAAI Conference on Artificial Intelligence; 2019 Jan 27–Feb 1; Honolulu, HI, USA; 2019.
[365] Kleiman-Weiner M, Shaw A, Tenenbaum JB. Constructing social preferences from anticipated judgments: when impartial inequity is fair and why? In: Proceedings of the 2017 Annual Meeting of the Cognitive Science Society; 2017 Jul 26–29; London, UK; 2017.
[366] Kleiman-Weiner M, Saxe R, Tenenbaum JB. Learning a commonsense moral theory. Cognition 2017;167:107–23.
[367] Kinney M, Tsatsoulis C. Learning communication strategies in multiagent systems. Appl Intell 1998;9(1):71–91.
[368] Lowe R, Wu Y, Tamar A, Harb J, Abbeel OP, Mordatch I. Multi-agent actor-critic for mixed cooperative–competitive environments. In: Proceedings of the 2017 Neural Information Processing Systems; 2017 Dec 3–9; Long Beach, CA, USA; 2017.
[369] Foerster J, Assael IA, de Freitas N, Whiteson S. Learning to communicate with deep multi-agent reinforcement learning. Proceedings of the 2016 Neural Information Processing Systems; 2016 Dec 5–10; Barcelona, Spain, 2016.
[370] Foerster J, Nardelli N, Farquhar G, Afouras T, Torr PHS, Kohli P, et al. Stabilising experience replay for deep multi-agent reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia. p. 1146–55.
[371] Holyoak KJ. Analogy and relational reasoning. In: Holyoak KJ, Morrison RG, editors. The Oxford handbook of thinking and reasoning. New York: Oxford University Press; 2012. p. 234–59.
[372] Raven JC. Raven progressive matrices. Torrance: Western Psychological Services; 1938.
[373] Zhang C, Gao F, Jia B, Zhu Y, Zhu SC. RAVEN: a dataset for relational and analogical visual reasoning. In: Proceedings of the 2019 Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA. p. 5317–27.
[374] Legg S, Hutter M. Universal intelligence: a definition of machine intelligence. Minds Mach 2007;17(4):391–444.
[375] Mo K, Zhu S, Chang AX, Yi L, Tripathi S, Guibas LJ, et al. PartNet: a large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: Proceedings of the 2019 Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA. p. 909–18.
[376] Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, et al. ShapeNet: an information-rich 3D model repository. 2015. arXiv:1512.03012.
[377] Feng T, Yu LF, Yeung SK, Yin K, Zhou K. Crowd-driven mid-scale layout design. ACM Trans Graph 2016;35(4):132.
[378] Savva M, Chang AX, Dosovitskiy A, Funkhouser T, Koltun V. MINOS: multimodal indoor simulator for navigation in complex environments. 2017. arXiv:1712.03931.
[379] Brodeur S, Perez E, Anand A, Golemo F, Celotti L, Strub F, et al. HoME: a household multimodal environment. 2017. arXiv:1711.11017.
[380] Xia F, Zamir AR, He Z, Sax A, Malik J, Savarese S. Gibson Env: real-world perception for embodied agents. In: Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–22; Salt Lake City, UT, USA; 2018. p. 9068–79.
[381] Wu Y, Wu YX, Gkioxari G, Tian Y. Building generalizable agents with a realistic and rich 3D environment. 2018. arXiv:1801.02209.
[382] Kolve E, Mottaghi R, Han W, VanderBilt E, Weihs L, Herrasti A, et al. AI2-THOR: an interactive 3D environment for visual AI. 2017. arXiv:1712.05474.
[383] Puig X, Ra K, Boben M, Li J, Wang T, Fidler S, et al. VirtualHome: simulating household activities via programs. In: Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–22; Salt Lake City, UT, USA; 2018. p. 8494–502.
[384] Xie X, Liu H, Zhang Z, Qiu Y, Gao F, Qi S, et al. VRGym: a virtual testbed for physical and interactive AI. In: Proceedings of the ACM TURC; 2019 May 17–19; Chengdu, China; 2019.

[385] Gao X, Gong R, Shu T, Xie X, Wang S, Zhu SC. VRKitchen: an interactive 3D virtual environment for task-oriented learning. 2019. arXiv:1903.05757.
[386] Shah S, Dey D, Lovett C, Kapoor A. AirSim: high-fidelity visual and physical simulation for autonomous vehicles. In: Hutter M, Siegwart R, editors. Field and service robotics. Cham: Springer; 2018. p. 621–35.
[387] Gao M, Wang X, Wu K, Pradhana A, Sifakis E, Yuksel C, et al. GPU optimization of material point methods. ACM Trans Graph 2018;37(6):254.
[388] Terzopoulos D, Platt J, Barr A, Fleischer K. Elastically deformable models. In: Stone MC, editor. Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques; 1987 July 27–31; Anaheim, CA, USA. New York: Association for Computing Machinery; 1987. p. 205–14.
[389] Terzopoulos D, Fleischer K. Modeling inelastic deformation: viscolelasticity, plasticity, fracture. In: Beach RJ, editor. Proceedings of the 15th Annual Conference on Computer Graphics and Interactive Techniques; 1988 Aug 1–5; Atlanta, GA, USA; New York: Association for Computing Machinery; 1988. p. 269–78.
[390] Foster N, Metaxas D. Realistic animation of liquids. Graph Models Image Proc 1996;58(5):471–83.
[391] Stam J. Stable fluids. ACM Trans Graph 1999;99:121–8.
[392] Bridson R. Fluid simulation for computer graphics. London: CRC Press; 2015.
[393] Bonet J, Wood RD. Nonlinear continuum mechanics for finite element analysis. New York: Cambridge University Press; 1997.
[394] Blemker S, Teran J, Sifakis E, Fedkiw R, Delp S. Fast 3D muscle simulations using a new quasistatic invertible finite-element algorithm. In: Proceedings of the 2005 International Symposium on Computer Simulation in Biomechanics; 2005 Jul 28–30; Cleveland, OH, USA; 2005.
[395] Hegemann J, Jiang C, Schroeder C, Teran JM. A level set method for ductile fracture. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation; 2013 Jul 19–21; Anaheim, CA, USA; 2013. p. 193–201.
[396] Gast TF, Schroeder C, Stomakhin A, Jiang C, Teran JM. Optimization integrator for large time steps. IEEE Trans Vis Comput Graph 2015;21(10):1103–15.
[397] Li M, Gao M, Langlois T, Jiang C, Kaufman DM. Decomposed optimization time integrator for large-step elastodynamics. ACM Trans Graph 2019;38(4):70.
[398] Wang Y, Jiang C, Schroeder C, Teran J. An adaptive virtual node algorithm with robust mesh cutting. In: Proceedings of the 2014 ACM SIGGRAPH/Eurographics Symposium on Computer Animation; 2014 Jul 21–23; Copenhagen, Denmark; 2014. p. 77–85.
[399] Monaghan JJ. Smoothed particle hydrodynamics. Annu Rev Astron Astrophys 1992;30(1):543–74.
[400] Liu WK, Jun S, Zhang YF. Reproducing kernel particle methods. Int J Numer Methods Fluids 1995;20(8–9):1081–106.
[401] Li S, Liu WK. Meshfree and particle methods and their applications. Appl Mech Rev 2002;55(1):1–34.
[402] Donea J, Giuliani S, Halleux JP. An arbitrary Lagrangian-Eulerian finite element method for transient dynamic fluid–structure interactions. Comput Methods Appl Mech Eng 1982;33(1–3):689–723.
[403] Brackbill JU, Ruppel HM. FLIP: a method for adaptively zoned, particle-in-cell calculations of fluid flows in two dimensions. J Comput Phys 1986;65(2):314–43.
[404] Jiang C, Schroeder C, Selle A, Teran J, Stomakhin A. The affine particle-in-cell method. ACM Trans Graph 2015;34(4):51.
[405] Sulsky D, Chen Z, Schreyer HL. A particle method for history-dependent materials. Comput Methods Appl Mech Eng 1994;118(1–2):179–96.
[406] Sulsky D, Zhou SJ, Schreyer HL. Application of a particle-in-cell method to solid mechanics. Comput Phys Commun 1995;87(1–2):236–52.
[407] Stomakhin A, Schroeder C, Chai L, Teran J, Selle A. A material point method for snow simulation. ACM Trans Graph 2013;32(4):102.
[408] Gaume J, Gast T, Teran J, van Herwijnen A, Jiang C. Dynamic anticrack propagation in snow. Nat Commun 2018;9(1):3047.
[409] Ram D, Gast T, Jiang C, Schroeder C, Stomakhin A, Teran J, et al. A material point method for viscoelastic fluids, foams and sponges. In: Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation; 2015 Aug 7–9; Los Angeles, CA, USA; 2015. p. 157–63.
[410] Yue Y, Smith B, Batty C, Zheng C, Grinspun E. Continuum foam: a material point method for shear-dependent flows. ACM Trans Graph 2015;34(5):160.
[411] Fang Y, Li M, Gao M, Jiang C. Silly rubber: an implicit material point method for simulating non-equilibrated viscoelastic and elastoplastic solids. ACM Trans Graph 2019;38(4):118.
[412] Klár G, Gast T, Pradhana A, Fu C, Schroeder C, Jiang C, et al. Drucker-Prager elastoplasticity for sand animation. ACM Trans Graph 2016;35(4):103.
[413] Daviet G, Bertails-Descoubes F. A semi-implicit material point method for the continuum simulation of granular materials. ACM Trans Graph 2016;35(4):102.
[414] Hu Y, Fang Y, Ge Z, Qu Z, Zhu Y, Pradhana A, et al. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. ACM Trans Graph 2018;37(4):150.
[415] Wang S, Ding M, Gast TF, Zhu L, Gagniere S, Jiang C, et al. Simulation and visualization of ductile fracture with the material point method. ACM Trans Graph 2019;2(2):18.
[416] Wolper J, Fang Y, Li M, Lu J, Gao M, Jiang C. CD-MPM: continuum damage material point methods for dynamic fracture animation. ACM Trans Graph 2019;38(4):119.
[417] Jiang C, Gast T, Teran J. Anisotropic elastoplasticity for cloth, knit and hair frictional contact. ACM Trans Graph 2017;36(4):152.
[418] Han X, Gast TF, Guo Q, Wang S, Jiang C, Teran J. A hybrid material point method for frictional contact with diverse materials. ACM Trans Graph 2019;2(2):17.
[419] Fu C, Guo Q, Gast T, Jiang C, Teran J. A polynomial particle-in-cell method. ACM Trans Graph 2017;36(6):222.
[420] Stomakhin A, Schroeder C, Jiang C, Chai L, Teran J, Selle A. Augmented MPM for phase-change and varied materials. ACM Trans Graph 2014;33(4):138.
[421] Tampubolon AP, Gast T, Klár G, Fu C, Teran J, Jiang C, et al. Multi-species simulation of porous sand and water mixtures. ACM Trans Graph 2017;36(4):105.
[422] Gao M, Pradhana A, Han X, Guo Q, Kot G, Sifakis E, et al. Animating fluid sediment mixture in particle-laden flows. ACM Trans Graph 2018;37(4):149.
[423] Nairn JA. Material point method calculations with explicit cracks. Comput Model Eng Sci 2003;4(6):649–64.
[424] Chen Z, Shen L, Mai YW, Shen YG. A bifurcation-based decohesion model for simulating the transition from localization to decohesion with the MPM. Z Angew Math Phys 2005;56(5):908–30.
[425] Schreyer HL, Sulsky DL, Zhou SJ. Modeling delamination as a strong discontinuity with the material point method. Comput Methods Appl Mech Eng 2002;191(23–24):2483–507.
[426] Sulsky D, Schreyer HL. Axisymmetric form of the material point method with applications to upsetting and Taylor impact problems. Comput Methods Appl Mech Eng 1996;139(1–4):409–29.
[427] Huang P, Zhang X, Ma S, Wang HK. Shared memory OpenMP parallelization of explicit MPM and its application to hypervelocity impact. Comput Model Eng Sci 2008;38(2):119–48.
[428] Hu W, Chen Z. Model-based simulation of the synergistic effects of blast and fragmentation on a concrete wall using the MPM. Int J Impact Eng 2006;32(12):2066–96.
[429] York AR, Sulsky D, Schreyer HL. Fluid-membrane interaction based on the material point method. Int J Numer Methods Eng 2000;48(6):901–24.
[430] Bandara S, Soga K. Coupling of soil deformation and pore fluid flow using material point method. Comput Geotech 2015;63:199–214.
[431] Guilkey JE, Hoying JB, Weiss JA. Computational modeling of multicellular constructs with the material point method. J Biomech 2006;39(11):2074–86.
[432] Huang P. Material point method for metal and soil impact dynamics problems. Beijing: Tsinghua University; 2010.
[433] Fang Y, Hu Y, Hu SM, Jiang C. A temporally adaptive material point method with regional time stepping. Comput Graph Forum 2018;37(8):195–204.
[434] Bardenhagen SG, Kober EM. The generalized interpolation material point method. Comput Model Eng Sci 2004;5(6):477–96.
[435] Gao M, Tampubolon AP, Jiang C, Sifakis E. An adaptive generalized interpolation material point method for simulating elastoplastic materials. ACM Trans Graph 2017;36(6):223.
[436] Sadeghirad A, Brannon RM, Burghardt J. A convected particle domain interpolation technique to extend applicability of the material point method for problems involving massive deformations. Int J Numer Methods Eng 2011;86(12):1435–56.
[437] Zhang DZ, Ma X, Giguere PT. Material point method enhanced by modified gradient of shape function. J Comput Phys 2011;230(16):6379–98.
[438] Bernstein DS, Givan R, Immerman N, Zilberstein S. The complexity of decentralized control of Markov decision processes. Math Oper Res 2002;27(4):819–40.
[439] Goldman CV, Zilberstein S. Optimizing information exchange in cooperative multi-agent systems. In: Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems; 2003 Jul 14–18; Melbourne, VIC, Australia. p. 137–44.
[440] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. 2013. arXiv:1312.5602.
[441] Tampuu A, Matiisen T, Kodelja D, Kuzovkin I, Korjus K, Aru J, et al. Multiagent cooperation and competition with deep reinforcement learning. PLoS ONE 2017;12(4):e0172395.
[442] Foerster JN, Farquhar G, Afouras T, Nardelli N, Whiteson S. Counterfactual multi-agent policy gradients. In: Proceedings of the 2018 AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA; 2018.
[443] Sukhbaatar S, Fergus R. Learning multiagent communication with backpropagation. In: Proceedings of the 2016 Neural Information Processing Systems; 2016 Dec 5–10; Barcelona, Spain; 2016. p. 2244–52.
[444] Mordatch I, Abbeel P. Emergence of grounded compositional language in multi-agent populations. In: Proceedings of the 2018 AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA; 2018.
[445] Lazaridou A, Peysakhovich A, Baroni M. Multi-agent cooperation and the emergence of (natural) language. In: Proceedings of the 5th International Conference on Learning Representations; 2017 Apr 24–26; Toulon, France; 2017.
[446] Havrylov S, Titov I. Emergence of language with multi-agent games: learning to communicate with sequences of symbols. In: Proceedings of the 2017 Neural Information Processing Systems; 2017 Dec 3–9; Long Beach, CA, USA; 2017.
[447] Evtimova K, Drozdov A, Kiela D, Cho K. Emergent language in a multi-modal, multi-step referential game. 2017. arXiv:1705.10369.
[448] Lazaridou A, Hermann KM, Tuyls K, Clark S. Emergence of linguistic communication from referential games with symbolic and pixel input. In: Proceedings of the 2018 International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada; 2018.

[449] Wagner K, Reggia JA, Uriagereka J, Wilkinson GS. Progress in the simulation of emergent communication and language. Adapt Behav 2003;11(1):37–69.
[450] Ibsen-Jensen R, Tkadlec J, Chatterjee K, Nowak MA. Language acquisition with communication between learners. J R Soc Interface 2018;15(140):20180073.
[451] Graesser L, Cho K, Kiela D. Emergent linguistic phenomena in multi-agent communication games. 2019. arXiv:1901.08706.
[452] Dupoux E, Jacob P. Universal moral grammar: a critical appraisal. Trends Cogn Sci 2007;11(9):373–8.
[453] Mikhail J. Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment. New York: Cambridge University Press; 2011.
[454] Blake PR, McAuliffe K, Corbit J, Callaghan TC, Barry O, Bowie A, et al. The ontogeny of fairness in seven societies. Nature 2015;528(7581):258–61.
[455] Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, et al. In search of homo economicus: behavioral experiments in 15 small-scale societies. Am Econ Rev 2001;91(2):73–8.
[456] House BR, Silk JB, Henrich J, Barrett HC, Scelza BA, Boyette AH, et al. Ontogeny of prosocial behavior across diverse societies. Proc Natl Acad Sci USA 2013;110(36):14586–91.
[457] Graham J, Meindl P, Beall E, Johnson KM, Zhang L. Cultural differences in moral judgment and behavior, across and within societies. Curr Opin Psychol 2016;8:125–30.
[458] Hurka T. Virtue, vice, and value. Cambridge: Oxford University Press; 2000.
[459] Rawls J. A theory of justice. Cambridge: Harvard University Press; 1971.
[460] Haidt J. The new synthesis in moral psychology. Science 2007;316 (5827):998–1002.
[461] Hamlin JK. Moral judgment and action in preverbal infants and toddlers: evidence for an innate moral core. Curr Dir Psychol Sci 2013;22 (3):186–93.
[462] Kim R, Kleiman-Weiner M, Abeliuk A, Awad E, Dsouza S, Tenenbaum JB, et al. A computational model of commonsense moral decision making. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society; 2018 Feb 2–3; New Orleans, LA, USA; 2018. p. 197–203.
[463] Holyoak KJ, Thagard P. The analogical mind. Am Psychol 1997;52(1): 35–44.
[464] Buehner MJ, Cheng PW. Causal learning. In: Holyoak KJ, Morrison RG, editors. The Oxford handbook of thinking and reasoning. New York: Oxford University Press; 2012. p. 210–33.
[465] Hesse MB. Models and analogies in science. South Bend: Notre Dame University Press; 1966.
[466] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 2013 Neural Information Processing Systems; 2013 Dec 5–8; Lake Tahoe, NV, USA; 2013.
[467] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013. arXiv:1301.3781.
[468] Carpenter PA, Just MA, Shell P. What one intelligence test measures: a theoretical account of the processing in the Raven progressive matrices test. Psychol Rev 1990;97(3):404–31.
[469] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, et al. VQA: visual question answering. In: Proceedings of the 2015 International Conference on Computer Vision; 2015 Dec 11–18; Santiago, Chile; 2015. p. 2425–33.
[470] Snow RE, Kyllonen PC, Marshalek B. The topography of ability and learning correlations. Adv Psychol Hum Intell 1984;2(S47):103.
[471] Jaeggi SM, Buschkuehl M, Jonides J, Perrig WJ. Improving fluid intelligence with training on working memory. Proc Natl Acad Sci USA 2008;105 (19):6829–33.
[472] Bower GH. A contrast effect in differential conditioning. J Exp Psychol 1961;62(2):196–9.
[473] Meyer DR. The effects of differential rewards on discrimination reversal learning by monkeys. J Exp Psychol 1951;41(4):268–74.
[474] Schrier AM, Harlow HF. Effect of amount of incentive on discrimination learning by monkeys. J Comp Physiol Psychol 1956;49(2):117–22.
[475] Shapley RM, Victor JD. The effect of contrast on the transfer properties of cat retinal ganglion cells. J Physiol 1978;285(1):275–98.
[476] Lawson R. Brightness discrimination performance and secondary reward strength as a function of primary reward amount. J Comp Physiol Psychol 1957;50(1):35–9.
[477] Amsel A. Frustrative nonreward in partial reinforcement and discrimination learning: some recent history and a theoretical extension. Psychol Rev 1962;69(4):306–28.
[478] Gibson JJ, Gibson EJ. Perceptual learning; differentiation or enrichment? Psychol Rev 1955;62(1):32–41.
[479] Gibson JJ. The ecological approach to visual perception: classic edition. London: Psychology Press; 2014.
[480] Catrambone R, Holyoak KJ. Overcoming contextual limitations on problem-solving transfer. J Exp Psychol Learn Mem Cogn 1989;15(6):1147–56.
[481] Gentner D, Gunn V. Structural alignment facilitates the noticing of differences. Mem Cognit 2001;29(4):565–77.
[482] Hammer R, Diesendruck G, Weinshall D, Hochstein S. The development of category learning strategies: what makes the difference? Cognition 2009;112 (1):105–19.
[483] Gick ML, Paterson K. Do contrasting examples facilitate schema acquisition and analogical transfer? Can J Psychol 1992;46(4):539.
[484] Haryu E, Imai M, Okada H. Object similarity bootstraps young children to action-based verb extension. Child Dev 2011;82(2):674–86.
[485] Smith L, Gentner D. The role of difference–detection in learning contrastive categories. In: Proceedings of the 2014 Annual Meeting of the Cognitive Science Society; 2014 Jul 23–26; Quebec City, QC, Canada; 2014.
[486] Gentner D. Structure-mapping: a theoretical framework for analogy. Cogn Sci 1983;7(2):155–70.
[487] Gentner D, Markman AB. Structural alignment in comparison: no difference without similarity. Psychol Sci 1994;5(3):152–8.
[488] Schwartz DL, Chase CC, Oppezzo MA, Chin DB. Practicing versus inventing with contrasting cases: the effects of telling first on learning and transfer. J Educ Psychol 2011;103(4):759–75.
[489] Zhang C, Jia B, Gao F, Zhu Y, Lu H, Zhu SC. Learning perceptual inference by contrasting. In: Proceedings of the 2019 Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada; 2019.
[490] Dehaene S. The number sense: how the mind creates mathematics. New York: Oxford University Press; 2011.
[491] Zhang W, Zhang C, Zhu Y, Zhu SC. Machine number sense: a dataset of visual arithmetic problems for abstract and relational reasoning. In: Proceedings of the 2020 AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York, NY, USA; 2020.