AI, Visual Imagery, and a Case Study on the Challenges Posed by Human Intelligence Tests

Maithilee Kunda[a,2]

This manuscript was compiled on January 14, 2021

Observations abound about the power of visual imagery in human intelligence, from how Nobel-prize-winning physicists make their discoveries to how children understand bedtime stories. These observations raise an important question for cognitive science, which is: what are the computations taking place in someone's mind when they use visual imagery? Answering this question is not easy and will require much continued research across the multiple disciplines of cognitive science. Here, we focus on a related and more circumscribed question from the perspective of artificial intelligence: if you have an intelligent agent that uses visual-imagery-based knowledge representations and reasoning operations, then what kinds of problem solving might be possible, and how would such problem solving work? We highlight recent progress in AI towards answering these questions in the domain of visuospatial reasoning, looking at a case study of how imagery-based artificial agents can solve visuospatial intelligence tests. In particular, we first examine several variations of imagery-based knowledge representations and problem-solving strategies that are sufficient for solving problems from the Raven's Progressive Matrices intelligence test. We then look at how artificial agents, instead of being designed manually by AI researchers, might learn portions of their own knowledge and reasoning procedures from experience, including learning visuospatial domain knowledge, learning and generalizing problem-solving strategies, and learning the actual definition of the task in the first place.

Artificial intelligence | Computational modeling | Mental imagery | Raven's Progressive Matrices | Visuospatial reasoning

> *"I think in pictures. Words are like a second language to me. I translate both spoken and written words into full-color movies, complete with sound, which run like a VCR tape in my head.... Language-based thinkers often find this phenomenon difficult to understand, but in my job as an equipment designer for the livestock industry, visual thinking is a tremendous advantage."*
> - Temple Grandin, prof. animal science and autism advocate (1)

> *"What I am really trying to do is bring birth to clarity, which is really a half-assedly thought-out pictorial semi-vision thing. I would see the jiggle-jiggle-jiggle or the wiggle of the path. Even now when I talk about the influence functional, I see the coupling and I take this turn–like as if there was a big bag of stuff–and try to collect it away and to push it. It's all visual. It's hard to explain."* - Richard Feynman, Nobel laureate in physics (2)

Temple Grandin is a well-known animal scientist who is on the autism spectrum. She has had incredible professional success in the livestock industry, and she credits her success to her strong visual imagery skills, i.e., abilities to generate, transform, combine, and inspect visual mental representations. (1).

Many physicists such as Richard Feynman (2), Albert Einstein (3) and James Clerk Maxwell (4) used imagery in their creative discovery processes, and similar patterns emerge in accounts by and about mathematicians (5), engineers (6), computer programmers (7), product designers (8), surgeons (9), memory champions (10), and more. People also use visual imagery in everyday activities such as language comprehension (11), story understanding (12), and physical (13) and mathematical reasoning (14).

These observations raise an interesting scientific question: what are the computations taking place in someone's mind when they use visual imagery? This is a difficult question that continues to receive attention across cognitive science disciplines (15).

Here, we focus on a related, more circumscribed question from the perspective of artificial intelligence: **IF you have an intelligent agent that uses visual-imagery-based knowledge representations and reasoning operations, THEN what kinds of problem solving might be possible, and how would it all work?**

In this paper, we discuss progress in AI towards answering this question in the domain of visuospatial reasoning—reasoning about the geometric and spatial properties of visual objects (16). This discussion necessarily leaves out such intriguing and important complexities as: non-visual forms of spatial reasoning, e.g., in people with visual impairments (17); the role of physics and forces in imagery (18); imagery in other sensory modalities (19); etc.

As a case study, we focus on visuospatial reasoning for solving human intelligence tests like Raven's Progressive Matrices. While many AI techniques have been developed to solve many different tests (20), we are still quite far from having an artificial agent that can "sit down and take" an intelligence test without specialized algorithms having been designed for that purpose. Contributions of this paper include discussions of:

1. Why intelligence tests are such a good challenge for AI.
2. A framework for artificial problem-solving agents with four components: a problem definition; input processing; domain knowledge; and a problem-solving strategy or procedure.
3. Several imagery-based agents that solve Raven's problems.
4. How an imagery-based agent could learn its domain knowledge, problem-solving strategies, and problem definition / input processing components, instead of each being manually designed.

[a] Electrical Engineering and Computer Science, Vanderbilt University, PMB 351679, 2301 Vanderbilt Place, Nashville, TN 37235-1679, USA
[2] To whom correspondence should be addressed. E-mail: mkunda@vanderbilt.edu

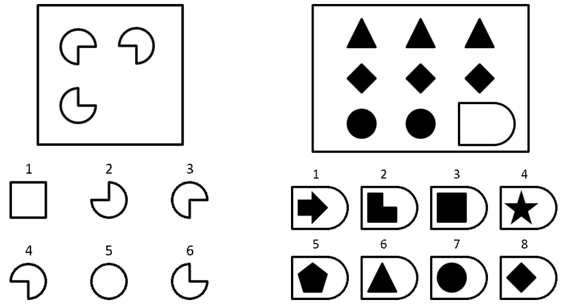

Fig. 1. Sample problems like those from the Raven's intelligence test, comparable to ones of easy-to-middling difficulty on the standard version of the test.

Why the Raven's test is (still!) a hard AI challenge

Take a look at the problems in Figure 1. Can you solve them?

While these problems may seem straightforward, consider for a moment the complexity of what you just did. As you were solving each problem, some executive control system in your mind was planning and executing a series of physical and cognitive operations, including shifts of gaze from one element of the problem to another; storing extracted features in working memory; computing and storing the results of intermediate calculations; and so on. And, you did all of this without any explicit instructions as to what cognitive operations to use, or in what order to apply them.

At a deeper level, you may notice that no one actually even told you what these problems were about. Typically, Raven's test-takers are instructed to solve each problem by selecting the answer from the bottom that best completes the matrix portion on top (21). However, even if you hadn't seen problems quite like these before, it is likely that you were able to grok the point of the problems just by looking at them, no doubt due to a lifetime of experience with pattern-matching games and multiple choice tests.

From a general AI perspective, intelligence tests like the Raven's have been "solved" in the sense that we do have computational programs that, given a Raven's problem as input, can often produce the correct answer as an output. In fact, some of the earliest work in AI was Evans' classic ANALOGY program from the 1960s—at the time, the largest program written in LISP to date!—that solved geometric analogy problems from college aptitude tests (22).

However, **all** of these programs have essentially been hand-crafted to solve Raven's problems in one way or another. Humans (at least in theory) are supposed to take intelligence tests without having practiced them beforehand. Thus, intelligence tests like the Raven's are still an "unsolved" challenge for AI when treated as tests of generalization, i.e., generalizing previously learned knowledge and skills to solve new and unfamiliar types of problems.

At an even higher level, the notion of "taking a test" is itself a sophisticated social and cultural construct. In people, for example, crucial research on stereotype threat has observed how stereotypes about race and gender can influence a person's performance on the exact same test depending on whether they are told it is a "test" or a "puzzle" (23). If we assume that human cognition can be explained in computational terms, then someday we ought to be able to have AI agents that model these effects.[*]

The Raven's test and similar tests of matrix reasoning and geometric analogy are particularly interesting for AI for several reasons. First, the Raven's test, originally designed to measure *eductive ability* or the ability to extract and understand information from a complex situation (21), occupies a unique niche among psychometric instruments as being the best single-format measure of a person's general intelligence (25). In other words, the Raven's test seems to tap into fundamental cognitive abilities that are very relevant to many other things a person tries to do.

Second, there are several Raven's tests that span a very wide range of difficulty levels, from problems that are easy for young children to problems that are difficult for most adults. The developmental trajectories of performance that people show offer a motivating parallel for studying AI agents that meaningfully improve their problem solving abilities through various learning experiences.

Third, there is evidence that many people use multiple forms of mental representation while solving Raven's problems, including inner language as well as visual imagery (26, 27). Interestingly, many people on the autism spectrum show patterns of performance on the Raven's test that do not match patterns seen in neurotypical individuals (28), and neuroimaging findings suggest that many individuals on the spectrum rely more on visual brain regions than neurotypicals do while solving the test (29). Thus, the Raven's test is a fascinating testbed for AI research on visual imagery in particular and multimodal reasoning more generally.

A framework for artificial agents that solve problems

Many approaches in AI can usefully be decomposed according to the framework shown in Figure 2. The agent is given a problem as input and is expected to produce a correct solution as output.

The **problem definition** refers to the agent's understanding of what the problem is actually asking, i.e., what constitutes a valid format of inputs and outputs (**problem template**) and what the goal is in terms of desired outputs (**solution criteria**). For example, for a generic Raven's problem, the problem template might specify a 2D matrix M of images m_i, with one entry in the matrix missing, and an unordered set A of answer images a_i, and that a valid answer consists of selecting one (and only one) answer $a_i \in A$. The solution criterion is that the selected answer should be the one that "best fits" in the missing slot in M.

The **input processing** component refers to how an agent takes raw or unstructured inputs from the "world" and converts them into a usable internal problem representation. For example, what the Raven's test actually provides is a pattern of ink on paper. At some point, this visual image needs to be decomposed into the matrix M and answer choice A elements in the problem template. For many artificial agents, input processing is performed outside the agent, either manually or by some other system. For example,

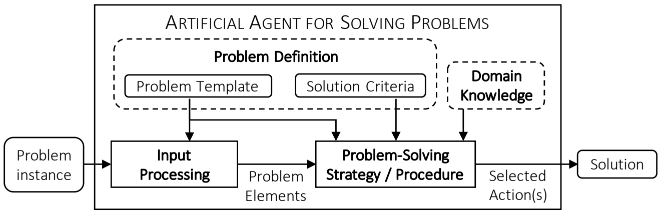

Fig. 2. Framework for artificial agents. Pushing the boundaries of what artificial agents can do often involves deriving more and more of the internal structure and knowledge of the agent through learning instead of programming.

[*] Perhaps ironically, early AI research studied what we thought were the hard problems, like taking tests and playing chess. The next wave of research recognized that the *real* hard problems were in fact the ones that were easy for many people, like walking around or recognizing cats (24). Now, we are realizing that the original hard problems of taking tests and playing chess are quite hard after all—but only if you really consider the full work of the agent, which includes figuring out what to do and understanding why you are doing this thing in the first place. In other words, many animals can walk around and pick up rocks, but only humans play good chess and take difficult tests.

most chess-playing agents do not operate using a video feed of a chess board, but rather using an explicit specification of where all the pieces are on the board. While this is a reasonable assumption to make in many AI applications, it does mean that the agent relies on having a simplified and pre-processed set of inputs.

Domain knowledge refers to whatever knowledge an agent needs to solve the given type of problems. The Raven's test can be tackled using visuospatial knowledge about symmetry, sequential geometric patterns, rows and columns, etc.

Finally, the **problem-solving strategy** encompasses what the agent actually does to solve a given problem, i.e., the algorithm that churns over the problem definition, domain knowledge, and specific problem inputs in order to generate an answer.

Given this framework, what would it mean for an agent to use visual imagery to solve problems? We offer one formulation: anywhere beyond the input processing step, the agent needs to use or retain representations of problem information that count as "images" in some way. This includes image-like representations occurring in the problem definition, domain knowledge, problem-solving strategy, and/or in the specific problem representations generated by the input processing component.

What counts as an image-like representation? Previous research on computational imagery often distinguishes between spatial representations, i.e., that replicate the spatial structure of what is being represented, versus visual/object representations, i.e., that replicate the visual appearance of what is being represented (30). These categories correspond to findings about spatial versus object imagery in people (31). Thus, we label agents using either type of representation as using visual imagery or being imagery-based. The imagery-based Raven's agents discussed later in this paper primarily use visual/object imagery and not spatial imagery, though certainly many other AI research efforts have developed agents that use spatial imagery (32).

Note that imagery here refers to the *format* in which something is represented, not the *contents* of what is represented. Many artificial agents reason about visuospatial information using non-imagery-based representations (33); for example, visuospatial domain knowledge can be encoded propositionally, such as the rule: `left-of(x,y)` $\implies$ `right-of(y,x)`.

Different types of Raven's problem-solving agents

Different paradigms of AI agents can now be described according to components in this framework.

Knowledge-based approaches, also associated with terms like cognitive systems (34) or symbolic AI, traditionally rely on manually designed domain knowledge and flexible problem-solving procedures like planning and search to tackle complex problems. The first wave of **propositional Raven's agents** used manual or automated input processing to convert raw test problem images into amodal, propositional representations, such as lists of attribute-value pairs, and then problem-solving procedures would operate over these propositional representations (33, 35–37). Visuospatial domain knowledge in these agents included predefined types of relationships among elements, like similarity or containment, and methods for extracting and defining relationships.

As foreshadowed in early writings about possible representational and algorithmic strategy differences on the Raven's test (38), a second wave of **imagery-based Raven's agents** were also knowledge-based but their internal representations of problem information remained visual, i.e., the problem-solving procedures directly accessed and manipulated problem images, and even often created new images during the course of reasoning (39–43). Visuospatial domain knowledge in these agents included image functions like rotation, image composition, visual similarity, etc.

More recently, a wave of **data-driven Raven's agents** aim to learn integrated representations of visuospatial domain knowledge and problem-solving strategies by training on input-output pairs from a large number of example problems (44–49).

Which approach is correct? This is a bad question, as different types of agents are used for very different lines of scientific inquiry. Referring again to Figure 2, most knowledge-based Raven's agents are used to study *problem-solving procedures* and assume a relatively fixed set of domain knowledge (though some of these agents certainly include forms of learning as well). Most of the data-driven Raven's agents are used to study how *domain knowledge* about visuospatial relationships can be learned from examples, and the problem-solving procedure is often (though not always) fixed.

All of these Raven's agents have many hand-built components, though the parts that are hand-built differ from one agent to another. Many open AI challenges remain, even within the one task domain of the Raven's test, in gradually converted the components in Figure 2 from being manually programmed to being learned or developed by the agents themselves. Next, we discuss how knowledge-based agents can use imagery to solve Raven's problems in several different ways, and then we examine emerging methods for agents to learn their own 1) domain knowledge, 2) problem-solving strategies, and finally 3) problem definitions.

Imagery-based strategies for solving Raven's problems

Within the category of imagery-based Raven's agents, many different formulations are possible, in terms of the problem-solving strategy that is used, the representation and contents of domain knowledge, and even the problem definition.

We describe five imagery-based strategies along with results from research by the author and colleagues. Results are reported for the Raven's Standard Progressive Matrices test, scored out of 60 problems (21). For comparison, human norm data suggests that average children in the US would score around 26/60 as 8-year-olds, 40/60 as 12-year-olds, and 49/60 as 16-yer-olds.

At a high level, the following strategies are described in terms of two strategy types observed in psychology research: (50):

- In *constructive matching*, the test-taker looks at the problem matrix, generates a guess for the missing element, and then chooses an answer most similar to its generated guess.
- In *response elimination*, the test-taker looks at each answer in turn, plugging it into the problem matrix, and choosing the one that produces the best overall matrix.

Strategy 1 (see Figure 3a). We developed an imagery-based agent that solves Raven's problems through multi-step search, using a constructive matching strategy (39, 43, 51):

1. Using elements from complete rows/columns of the matrix, search among known visual transformations for the one that best explains image variation across parallel rows/columns.
2. Apply this transformation to elements in a partial row or column to predict a new answer image.
3. Search among the answer choices to find the one that is most similar to the predicted answer image.

More formally, problem inputs include a set M of images m_i representing sections of the problem matrix, and a set A of answer choice images a_i. Let C be the set of all collinear subsets c of M,

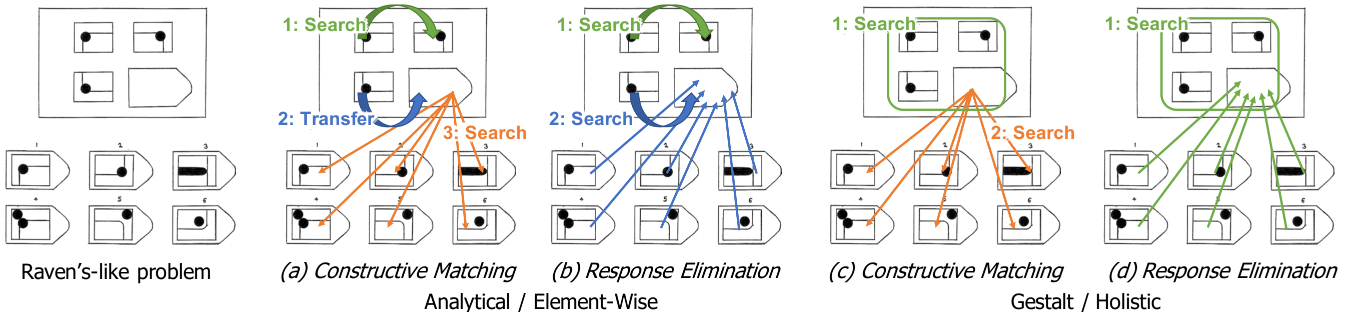

Fig. 3. Raven's-like problem and four different imagery-based strategies for solving it. A problem consists of matrix M of elements m_i and set A of answer choices a_i. **(a)** First strategy begins with search for transformation t that best transforms m_1 into m_2, then applies t to m_3 to produce an image candidate for m_4, and finally searches for answer a_i most similar to m_4. **(b)** Second strategy also begins with search for t that best transforms m_1 into m_2, then conducts similar searches for transformations t_{ai} that transform m_3 into each a_i, and finally searches for answer a_i that yields t_{ai} most similar to t. **(c)** Third strategy begins with search for image m_4 that maximizes Gestalt metric for matrix M, and then searches for answer a_i most similar to m_4. **(d)** Fourth strategy involves search for answer a_i that maximizes Gestalt metric for matrix M.

with c_x referring to the first element(s), and c_y referring to the last element. Each c contains matrix elements along rows, columns, or diagonals. We define an analogy g as a pairing of a single complete collinear subset c_1 with an incomplete collinear subset c_2 (i.e., $g = [c_{1.x} : c_{1.y} :: c_{2.x} : c_{2.y}]$, where $c_{2.y}$ is the missing element in the matrix). All such analogies that share the same c_2 are further aggregated into sets $G_i \in G$.

In addition, let T be the agent's predefined set of visual transformations. Also let $\mathrm{sim}(\mathrm{I}_1, \mathrm{I}_2)$ be a function that returns a real-valued measure of similarity between images I_1 and I_2. First, the agent finds the best-fit transformation:

$$(t_{max}, g_{max}) = \operatorname*{argmax}_{t \in T, G_i \in G} \left(\operatorname*{mean}_{g \in G_i} \left(\mathrm{sim}\left(t(g.c_{1.x}), g.c_{1.y} \right) \right) \right)$$

Second, the agent computes a predicted answer image as: $a_{pred} = t_{max}(g_{max}.c_{2.x})$. Third, the agent returns the most similar answer choice: $a_{final} = \mathrm{argmax}_{a_i \in A} \left(\mathrm{sim}(a_{pred}, a_i) \right)$.

Hand-coded domain knowledge is provided in the form of the set T of visual transformations, including eight rectilinear rotations and reflections (including identity) and three to six image composition operations (union, intersection, subtraction, and combinations of these) as well as visual similarity and other image processing utility functions. Steps 1 and 3 above used exhaustive search.

Successive versions of the agent, using more transformations T and more varied ways to optimize over matrix entries in Step 1, have achieved scores of 38/60 (39), 50/60 (51), and 57/60 (43) on the Raven's Standard Progressive Matrices test.

Strategy 2 (see Figure 3b). In a related line of research, colleagues developed a different imagery-based agent that adopted a response elimination type of strategy (see Figure 3b). In this work (40), a smaller set of visual transformations (rotation and reflection) was used to compute *fractal image transformations*, i.e. a representation of one image in terms of another, using techniques from image compression (52).

In particular, to compute a fractal transformation between source image A and target image B, B is first partitioned into a set of subimages b_i. Then, for each b_i, a fragment $a_i \in A$ is found such that b_i can be expressed as an affine transformation t_i of a_i. The fragments a_i are twice the size of b_i, resulting in a contractive transformations The set T of all t_i is the fractal transformation of A into B.

To solve a Raven's problem, a fractal transformation T is computed using elements from each complete row/column j in the matrix, and then similar transformations T'_{ij} are computed for each of the answer choices plugged into the incomplete rows/columns of the matrix. Finally, the selected answer is the one yielding the most similar fractal transformations to those computed for the original rows/columns of the matrix. Formally, if we let Tsim be a similarity metric across fractal transformations, the final answer is given by:

$$a_{final} = \operatorname*{argmax}_{a_i \in A} \sqrt{\sum_j \mathrm{Tsim}(T_j, T'_{ij})^2}$$

Results using this fractal method were also 50 out of 60 correct on the Raven's Standard Progressive Matrices test, allowing for some ambiguous detections of the answers, or 38 out of 60 correct with a specific method for resolving these ambiguities (40).

Strategy 3 (see Figure 3c). The first two strategies consider each matrix element individually. However, people can also use a "Gestalt" strategy to consider the entire matrix as a whole (38, 53). For instance, for the problem in Figure 3, if one looks at the matrix as a single image, an answer might just "appear" in the blank.

In recent work (42), we attempted to model this kind of strategy using neural networks for image inpainting, trained to fill in the missing portions of real photographs. We used a recently published image inpainting network consisting of a variational autoencoder combined with a generative adversarial network (54), and we tested several versions of the network trained on different types of photographs, such as objects, faces, scenes, and textures. Given an image of the incomplete problem matrix, the network outputs a guess for what image should fill in the missing portion. This guess is then used to select the most similar answer.

Formally, let F be the learned encoder network that converts an image into a representation in a learned feature space, and let G be the learned decoder network that converts a feature-based image back into pixel space, including inpainting to fill in any missing portions. Then, our agent first computes $M' = G(F(M))$ to obtain a new, filled-in matrix image, with m_x denoting the new, filled in portion of M'. Let $\mathrm{L2dist}$ represents the L2 norm of a vector in the learned feature space. Then, the final answer is:

$$a_{final} = \operatorname*{argmin}_{a_i \in A} \left(\mathrm{L2dist}\left(F(m_x) - F(a_i) \right) \right)$$

Figure 4 shows examples of inpainting results on several example problems, some of which are filled in more effectively than others. The best version of this agent, trained on photographs of

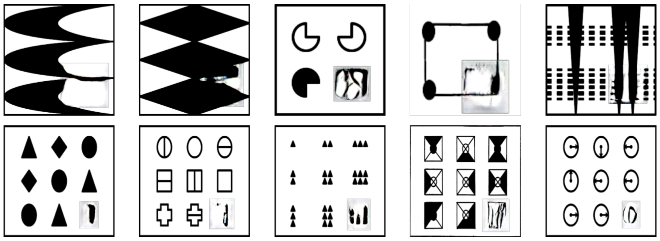

Fig. 4. Images generated using an inpainting neural network (54) for Raven's-like problems (42). The network was trained only on real-world photographs of objects.

objects, answered 25 out of 60 problems on the Raven's Standard Progressive Matrices test. While this score may seem low, it is quite astonishing given that there was no Raven's-specific information fed into or contained in the inpainting network, and in fact the network had never before "seen" line drawings, only photographs.

Strategy 4 (see Figure 3c). The fourth strategy combines a Gestalt approach with response elimination. We have not yet implemented this strategy, nor do we know of other AI efforts that have, but we present a brief sketch here. Essentially, this strategy works by plugging in answers to the matrix, and choosing the one that creates the "best" overall picture, for some notion of best.

Assume a Gestalt metric S that measures the Gestalt quality of any given image. Images that are highly symmetric, contain coherent objects, etc. would score highly, and images that are chaotic or broken up would score poorly. Then, the agent chooses the answer that scores highest when plugged into the matrix M:

$$a_{final} = \underset{a_i \in A}{\operatorname{argmax}} \left(S(M \cup a_i) \right)$$

Strategy 5 (not shown in figure). The above four strategies treat RPM matrix elements as single images. However, previous computational and human studies have suggested that it can be helpful to decompose RPM problems into multiple subproblems, by breaking up a single matrix element into subcomponents (35).

In previous work, we have also explored imagery-based techniques for decomposing a geometric analogy into subproblems, solving each separately, and then re-assembling the sub-solutions back together to choose the final answer (55), though this method has not yet been tested on the actual Raven's tests.

Open questions. From this small survey, it is clear that there is no single imagery-based Raven's strategy. Imagery-based agents are like logic-based agents or neural-network-based agents; there are a set of generally shared principles of representation and reasoning, but then individual agents are designed to use specific instantiations of these and combine them in different ways to produce very diverse problem-solving behaviors.

Exploring the space of imagery-based agents is valuable not to find the "best" one, but rather to characterize the space itself. Each agent, as a data point in this space of possible agents, is an artifact that can be studied in order to understand something about how *that* particular set of representations and strategies can produce intelligent task behaviors (56). Future work should continue to add data points to this space and also investigate the extent to which these strategies overlap with human problem-solving.

Learning visuospatial domain knowledge

Imagery-based agents use many kinds of visuospatial domain knowledge, including: visual transformations like rotation, scaling, and composition; hierarchical representations of concepts in terms of attributes like shape and texture; Gestalt principles like symmetry, continuity and similarity; etc. These types of knowledge can be leveraged by an agent to solve problems from the Raven's test as well as many other visuospatial tests (32).

Visuospatial domain knowledge also includes more semantically rich information such as what kinds of objects go where in a scene (57); we do not further discuss this type of semantic knowledge here, though it certainly plays an important role in imagery-based AI, especially for agents that perform language understanding or commonsense reasoning tasks (32).

How is visuospatial domain knowledge learned? One hypothesis suggests that agents learn such knowledge through prior sensorimotor interactions with the world. Under this view, the precise nature of the representations and learning mechanisms involved are important open questions. For brevity, we discuss here AI research on learning two types of visuospatial domain knowledge—visual transformations and Gestalt principles.

Learning visual transformations. In humans, many reasoning operators used during visual imagery (e.g., transformations like mental rotation, scaling, etc.) are hypothesized to be learned from visuomotor experience, e.g., perceiving the movement of physical objects in the real world (58). As with the well-known kittens-in-carousel experiments (59), learning visual transformations may rely on the combination of active motor actions coupled with visual perception of the results of those actions. Studies in both children and adults have indeed found that training on a manual rotation task does improve performance on mental rotation (60, 61).

Computational efforts to model the learning of visual transformations have generally represented each transformation as a set of weights in a neural network. In early work, distinct networks were used to learn each transformation individually (62). More recent work combines the visual and motor components of inputs for learning mental rotation (63). While many of these approaches implement visual transformations as distinct operations, a more general approach might represent continuous visual operations as combinations of basis functions that can be combined in arbitrary ways (64). Along these lines, other recent work uses more complex neural networks to represent transformations as combinations of multiple learned factors, though this work still focused on relatively simple transformations like rotation and scaling (65, 66).

People certainly do not learn visual transformations from specialized training on rotation, scaling, etc., taken as separate transformations. More generally, we have access to a very robust and diverse machinery for simulating visual change, and the simple "mental rotation" types of tasks often used in studies of visual imagery tap into only very tiny slices of this knowledge base. In line with evidence of the importance of motor actions and forces on our own imagery abilities (18), we expect that work in AI to model physical transformations—especially work in robotics that combines visual and motor inputs/outpus—will be essential for producing the kinds of capabilities agents need for visual imagery.

There is starting to be a wave of relevant work in AI in the area of *video prediction*, which involves learning representations of both the appearance of objects as well as their dynamics (67–69), including for increasingly complex forms of dynamics as with a robot trying to manipulate a rope (70). Importantly, these efforts focus

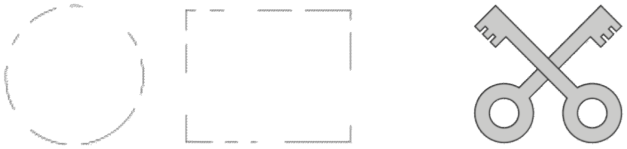

Fig. 5. Images eliciting Gestalt "completion" phenomena. Left contains only scattered line segments, but we inescapably see a circle and rectangle. Right contains one whole key and one broken key, but we see two whole keys with occlusion.

on learning and making inferences about object dynamics directly in the image space, as opposed to computational approaches that rely on explicit physics simulations and then project predictions into image space. Thus, these new approaches offer intriguing possibilities as potential models for how humans might learn naive physics as a form of imagery-based reasoning.

Learning Gestalt principles. Many visuospatial intelligence tests rely on a person's knowledge of visual relationships like similarity, continuity, symmetry, etc. Simple tests like shape matching require the test-taker to infer first-order relationships among visual elements, while more complex tests like the Raven's often progress into second-order relationships, i.e., relations over relations.

In one sense, a test like the Raven's ought to be agnostic with respect to the specific choice of first-order relationships, and indeed in many propositional AI agents, a relation like `contains(X,Y)` can be replaced with any arbitrary label, and the results will stay the same. However, for people, the actual visuospatial relationships at play do deeply influence our problem-solving capabilities. For example, isomorphs of the Tower of Hanoi task are more difficult if task rules are less well-aligned with our real-world knowledge about spatial structure and stacking (71). Similarly, the perceptual properties of Raven's problems have been found to be a strong predictor of item difficulty (72).

A person's prior knowledge about visuospatial relationships is closely tied to Gestalt perceptual phenomena. In humans, Gestalt phenomena have to do, in part, with how we integrate low-level perceptual elements into coherent, higher-level wholes (73), as shown in Figure 5. Psychology research has enumerated a list of principles (or laws, perceptual/reasoning processes, etc.) that seem to operate in human perception, like preferences for closure, symmetry, etc. (74). Likewise, work in image processing and computer vision has attempted to define these principles mathematically or computationally, for instance as a set of rules (75).

However, in more recent computational models, Gestalt principles are seen as emergent properties that reflect, rather than determine, perceptions of structure in an agent's visual environment. For example, early approaches to image inpainting—i.e., reconstructing a missing/degraded part of an image—used rule-like principles to determine the structure of missing content, while later approaches use machine learning to capture structural regularities from data and apply them to new images (76). This seems reasonable as a model of Gestalt phenomena in human cognition; it is our years of experience with the world around us we see Figure 5 (left) as partially occluded/degraded views of whole objects.

Image inpainting represents a fascinating area of imagery-based abilities for artificial agents (54), which we used in our model of Gestalt-type problem solving on the Raven's test (42), as described earlier. Other work in computer vision and machine learning studies the extent to which neural networks not explicitly designed to model Gestalt effects might exhibit such effects as emergent phenomena (77–81).

Learning a problem-solving strategy

Relatively little research in AI has proposed methods for automatically generating problem-solving procedures for intelligence tests, despite the extensive research on manually constructed solution methods or methods that rely on a large number of examples (20). How does a person obtain an effective problem-solving strategy for a task they have never seen, on the fly and often without explicit feedback? Some human research suggests that children learn to solve a widening range of problems through two primary processes of 1) *strategy discovery*, i.e., discovering new strategies for certain problems or tasks, and 2) *strategy generalization*, i.e., adapting strategies they already know for other problems or tasks (82, 83).

Some AI research on strategy discovery can be found in the area of inductive programming or program synthesis, i.e., given a number of input-output pairs, constraints, or other partial specifications of a task, together with a set of available operations, the system induces a "program" or series of operations that produces the desired behaviors (84). In other words, "Inductive programming can be seen as a very special subdomain of machine learning where the hypothesis space consists of classes of computer programs" (85). Inductive programming has been applied to some intelligence-test-like tasks, such as number series problems (86), and to simple visual tasks like learning visual concepts (87, 88). However, more research is needed to expand these methods to tackle more complex and diverse sets of tasks. For example, given the imagery-based strategies described above, a challenge for imagery-based program induction would be to derive these strategies automatically from a small set of example Raven's problems.

AI research has often investigated strategy generalization through the lens of integrating planning with analogy. Case-based planning looks at how plans stored in memory are retrieved at the appropriate juncture, modified, and applied to solve a new problem (89). The majority of this work has focused on agents that use propositional knowledge representations, and very little (if any) has applied these methods to address intelligence tests.

Research on strategy selection and adaptation would be enormously informative for studying not just how people approach a new type of intelligence test but also inter-problem learning on intelligence tests, i.e., learning from one problem (even without feedback) and use this knowledge to inform the solution of the next problem. In humans, one fascinating study gave each of two groups of children a different set of Raven's-like problems to start with, and then the same final set of problems that had ambiguous answers (53). Depending on which set of starting problems they received, the children predictably gravitated towards one of two profiles of performance on the final problems. Modeling these phenomena remains an open challenge for AI research.

Learning the problem definition

Even with intelligent agents that generate their own problem-solving strategies or programs, the problem definition—i.e., the problem template and goal—is still provided by the human system designer. Interactive task learning is an area of AI research that investigates how "an agent actively tries to learn the actual definition of a task through natural interaction with a human instructor, not just how to perform a task better" (90). Research in interactive task learning generally involves designing agents or robots that learn from both verbal and nonverbal information, i.e., instructions along with examples or situated experiences (91, 92).

Such multi-modal inputs are used all the time in human learning, including on intelligence tests: most tests combine verbal (spoken or written) instructions with simple example problems to teach the test-taker the point of each new task that is presented. For example, the Raven's test typically begins with spoken instructions to select the answer choice that best fills in the matrix, together with a very simple example problem that the test administrator is supposed to show the test-taker, along with the correct answer.

Any Raven's agent must contain information about the problem definition in order to parse new problems appropriately and to follow a procedure that attains the goal. Moreover, agents should be able to modify their problem definition to accommodate slight problem variations. For example, if a new problem is presented with two empty spots in the matrix, a robust agent should be able to infer that this problem requires two corresponding answer responses.

In all extant Raven's agents, knowledge of the problem definition is manually provided by system designers. While these concepts may seem straightforward to a person, and indeed are usually trivial to program into an agent as static program elements, it is a challenging open question to consider where these concepts come from, and how they might be learned. For example, people gain extensive experience in taking multiple choice tests from a very early age, especially in modern societies, but we do not know precisely how this knowledge is represented, or the mechanisms by which it is generalized to new tasks.

The interesting sub-problem of *nonverbal task learning* considers how the task definition can be learned purely through a small number of observed examples, without the use of explicit language-based information at all (93). While nonverbal mechanisms are undoubtedly at play in multi-modal task learning for most people, nonverbal task learning in its pure form does also occur.

There are many clinical populations in which individuals have difficulties in using or understanding language, including acquired aphasias or developmental language disorders. Nonverbal intelligence tests are specifically designed for use with such populations, and they avoid verbal instructions altogether (94). In these tests, examiners initially show test-takers a simple example problem and its solution. Test-takers must learn the task definition (e.g., matching shapes, finding one shape in another, completing a visual pattern, etc.) by observing the example, and then use this knowledge to solve a series of more difficult test problems.

A small but intriguing set of converging research threads in AI have pinpointed the importance of nonverbal task learning. One recent study using robots looked at how abstract goals can be inferred from a small number of visual problem examples and applied to new problems, where the goal is represented in terms of a set of programs that meets it (95). Even more recently, a new Abstraction and Reasoning Corpus (ARC) has been proposed for artificial agents, containing 1,000 visual tasks with distinct goals; agents must infer the goal for a given task from a few examples and then use this knowledge to solve new problems (96). Both of these tasks are similar to the Raven's test in the sense that, even though the Raven's test ostensibly only has a single goal (i.e. choose the answer that fits best), different Raven's problems can be thought of as requiring different formulations of this overarching and extremely vague goal. These examples also pose interesting questions about the extent to which problem goals might be implicitly represented within an agent's problem-solving strategy, instead of explicitly, and the pros and cons of each alternative.

Note that this discussion only considers goals that are well-defined at least in the minds of the problem creators. Intelligence tests are a rather odd social construct for this reason; in a way, the test-taker is trying to infer the intent of the test designer. How agents (or humans) represent and reason about their *own* goals might involve an extension of the processes described here, or they might be different modes of reasoning altogether.

Conclusion and implications for cognitive science

We close by returning to the motivating questions from the introduction. The cognitive science question is: what are the computations taking place in someone's mind when they use visual imagery?

AI research alone cannot, of course, fully answer this question, and so we presented a second, more limited question: if you have an intelligent agent that uses visual-imagery-based knowledge representations and reasoning operations, then what kinds of problem solving might be possible, and how would it all work?

In this paper, we have presented a review of AI research and open lines of inquiry related to answering this question in the context of imagery-based agents that solve problems from the Raven's Progressive Matrices intelligence test. We discussed: 1) why intelligence tests are such a good challenge for AI; 2) A framework for artificial problem-solving agents; 3) several imagery-based agents that solve Raven's problems; and 4) how an imagery-based agent could *learn* its domain knowledge, problem-solving strategies, and problem definition, instead of these components being manually designed and programmed.

More generally, whether or not imagery-based AI agents are at all similar to humans, designing, implementing, and studying such agents contributes valuable information about what is *possible* in terms of computation and intelligence. AI research that develops different kinds of agents is helpful for sketching out different points in the space of what is possible, and AI research that enables such agents to learn is helpful for hypothesizing how and why various computational elements of intelligence might come to be. Then, further interdisciplinary inquiries can proceed to connect findings and hypotheses derived from these lines of AI research to corresponding lines of research about what humans do.

Acknowledgments

Many thanks to the reviewers for their helpful comments. This work was funded in part by NSF award #1730044.

1 T Grandin, *Thinking in pictures, expanded edition: My life with autism.* (Vintage), (2008).

2 J Gleick, *Genius: The life and science of Richard Feynman.* (Vintage), (1992).

3 GJ Feist, *The psychology of science and the origins of the scientific mind.* (Yale University Press), (2008).

4 NJ Nersessian, *Creating scientific concepts.* (MIT press), (2008).

5 M Giaquinto, *Visual thinking in mathematics.* (Oxford University Press), (2007).

6 ES Ferguson, *Engineering and the Mind's Eye.* (MIT press), (1994).

7 M Petre, AF Blackwell, Mental imagery in program design and visual programming. *Int. J. Human-Computer Stud.* **51**, 7–30 (1999).

8 DW Dahl, A Chattopadhyay, GJ Gorn, The use of visual mental imagery in new product design. *J. Mark. Res.* **36**, 18–28 (1999).

9 KR Wanzel, SJ Hamstra, DJ Anastakis, ED Matsumoto, MD Cusimano, Effect of visual-spatial ability on learning of spatially-complex surgical skills. *The lancet* **359**, 230–231 (2002).

10 J Foer, *Moonwalking with Einstein: The art and science of remembering everything.* (Penguin), (2011).

11 BK Bergen, *Louder than words: The new science of how the mind makes meaning.* (Basic Books (AZ)), (2012).

12 JS Hutton, et al., Home reading environment and brain activation in preschool children listening to stories. *Pediatrics* **136**, 466–478 (2015).

13 M Hegarty, Mechanical reasoning by mental simulation. *Trends cognitive sciences* **8**, 280–285 (2004).

14 D Van Garderen, Spatial visualization, visual imagery, and mathematical problem solving of students with varying abilities. *J. learning disabilities* **39**, 496–506 (2006).

15 J Pearson, SM Kosslyn, The heterogeneity of mental representation: ending the imagery debate. *Proc. Natl. Acad. Sci.* **112**, 10089–10092 (2015).

16 NS Newcombe, TF Shipley, Thinking about spatial thinking: New typology, new assessments in *Studying visual and spatial reasoning for design creativity.* (Springer), pp. 179–192 (2015).

17 M Knauff, E May, Mental imagery, reasoning, and blindness. *Q. J. Exp. Psychol.* **59**, 161–177 (2006).

18 DL Schwartz, Physical imagery: Kinematic versus dynamic models. *Cogn. Psychol.* **38**, 433–464 (1999).

19 MO Belardinelli, et al., An fmri investigation on image generation in different sensory modalities: the influence of vividness. *Acta psychologica* **132**, 190–200 (2009).

20 J Hernández-Orallo, F Martínez-Plumed, U Schmid, M Siebers, DL Dowe, Computer models solving intelligence test problems: Progress and implications. *Artif. Intell.* **230**, 74–107 (2016).

21 J Raven, JC Raven, JH Court, *Manual for Raven's Progressive Matrices and Vocabulary Scales.* (Harcourt Assessment, Inc.), (1998).

22 TG Evans, A program for the solution of geometric-analogy intelligence test questions in *Semantic Information Processing*, ed. M Minsky. (MIT Press, Cambridge, MA), pp. 271–353 (1968).

23 RP Brown, EA Day, The difference isn't black and white: Stereotype threat and the race gap on raven's advanced progressive matrices. *J. Appl. Psychol.* **91**, 979 (2006).

24 RA Brooks, Intelligence without representation. *Artif. intelligence* **47**, 139–159 (1991).

25 RE Snow, PC Kyllonen, B Marshalek, The topography of ability and learning correlations. *Adv. psychology human intelligence* **2**, 47–103 (1984).

26 V Prabhakaran, JA Smith, JE Desmond, GH Glover, JD Gabrieli, Neural substrates of fluid reasoning: an fMRI study of neocortical activation during performance of the Raven's Progressive Matrices test. *Cogn. psychology* **33**, 43–63 (1997).

27 RP DeShon, D Chan, DA Weissbein, Verbal overshadowing effects on Raven's advanced progressive matrices: Evidence for multidimensional performance determinants. *Intelligence* **21**, 135–155 (1995).

28 M Dawson, I Soulières, MA Gernsbacher, L Mottron, The Level and Nature of Autistic Intelligence. *Psychol. Sci.* **18**, 657–662 (2007).

29 I Soulières, et al., Enhanced visual processing contributes to matrix reasoning in autism. *Hum. Brain Mapp.* **30**, 4082–4107 (2009).

30 J Glasgow, D Papadias, Computational Imagery. *Cogn. Sci.* **16**, 355–394 (1992).

31 M Kozhevnikov, S Kosslyn, J Shephard, Spatial versus object visualizers: A new characterization of visual cognitive style. *Mem. & cognition* **33**, 710–726 (2005).

32 M Kunda, Visual mental imagery: A view from artificial intelligence. *Cortex* **105**, 155–172 (2018).

33 A Lovett, K Forbus, Modeling visual problem solving as analogical reasoning. *Psychol. review* **124**, 60 (2017).

34 P Langley, The cognitive systems paradigm. *Adv. Cogn. Syst.* **1**, 3–13 (2012).

35 PA Carpenter, MA Just, P Shell, What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychol. review* **97**, 404–431 (1990).

36 D Rasmussen, C Eliasmith, A neural model of rule generation in inductive reasoning. *Top. Cogn. Sci.* **3**, 140–153 (2011).

37 C Strannegård, S Cirillo, V Ström, An anthropomorphic method for progressive matrix problems. *Cogn. Syst. Res.* **22**, 35–46 (2013).

38 E Hunt, Quote the Raven? Nevermore in *Knowledge and cognition.* (Lawrence Erlbaum, Oxford, England), pp. ix, 321 (1974).

39 M Kunda, K McGreggor, AK Goel, A computational model for solving problems from the raven's progressive matrices intelligence test using iconic visual representations. *Cogn. Syst. Res.* **22**, 47–66 (2013).

40 K McGreggor, M Kunda, AK Goel, Fractals and ravens. *Artif. Intell.* **215**, 1–23 (2014).

41 S Shegheva, A Goel, The structural affinity method for solving the raven's progressive matrices test for intelligence in *Thirty-Second AAAI Conference on Artificial Intelligence.* (2018).

42 T Hua, M Kunda, Modeling gestalt visual reasoning on raven's progressive matrices using generative image inpainting techniques in *Annual Conference on Advances in Cognitive Systems (ACS).* (2020).

43 Y Yang, K McGreggor, M Kunda, Not quite any way you slice it: How different analogical constructions affect raven's matrices performance in *Annual Conference on Advances in Cognitive Systems (ACS).* (2020).

44 D Hoshen, M Werman, Iq of neural networks (2017).

45 DG Barrett, F Hill, A Santoro, AS Morcos, T Lillicrap, Measuring abstract reasoning in neural networks (2018).

46 F Hill, A Santoro, DG Barrett, AS Morcos, T Lillicrap, Learning to make analogies by contrasting abstract relational structure (2019).

47 X Steenbrugge, S Leroux, T Verbelen, B Dhoedt, Improving generalization for abstract reasoning tasks using disentangled feature representations (2018).

48 S van Steenkiste, F Locatello, J Schmidhuber, O Bachem, Are disentangled representations helpful for abstract visual reasoning? (2019).

49 C Zhang, F Gao, B Jia, Y Zhu, SC Zhu, Raven: A dataset for relational and analogical visual reasoning in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* pp. 5317–5327 (2019).

50 CE Bethell-Fox, DF Lohman, RE Snow, Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence* **8**, 205–238 (1984).

51 M Kunda, Ph.D. thesis (Georgia Tech) (2013).

52 M Barnsley, LP Hurd, *Fractal Image Compression.* (A.K. Peters, Boston, MA), (1992).

53 JR Kirby, MJ Lawson, Effects of strategy training on progressive matrices performance. *Contemp. Educ. Psychol.* **8**, 127–140 (1983).

54 J Yu, et al., Generative image inpainting with contextual attention in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* pp. 5505–5514 (2018).

55 M Kunda, Computational mental imagery, and visual mechanisms for maintaining a goal-subgoal hierarchy in *Proceedings of the Third Annual Conference on Advances in Cognitive Systems (ACS).* p. 4 (2015).

56 A Newell, HA Simon, Computer science as empirical inquiry: Symbols and search. *Commun. ACM* **19**, 113–126 (1976).

57 AX Chang, M Savva, CD Manning, Learning spatial knowledge for text to 3d scene generation. in *EMNLP.* pp. 2028–2038 (2014).

58 RN Shepard, Ecological constraints on internal representation: resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychol. review* **91**, 417–447 (1984).

59 R Held, A Hein, Movement-produced stimulation in the development of visually guided behavior. *J. comparative physiological psychology* **56**, 872 (1963).

60 G Wiedenbauer, J Schmid, P Jansen-Osmann, Manual training of mental rotation. *Eur. J. Cogn. Psychol.* **19**, 17–36 (2007).

61 G Wiedenbauer, P Jansen-Osmann, Manual training of mental rotation in children. *Learn. instruction* **18**, 30–41 (2008).

62 BW Mel, A connectionist learning model for 3-d mental rotation, zoom, and pan in *Proceedings of the Eighth Annual Conference of the Cognitive Science Society.* pp. 562–71 (1986).

63 K Seepanomwan, D Caligiore, G Baldassarre, A Cangelosi, Modelling mental rotation in cognitive robots. *Adapt. Behav.* **21**, 299–312 (2013).

64 RP Goebel, The mathematics of mental rotations. *J. Math. Psychol.* **34**, 435–444 (1990).

65 R Memisevic, GE Hinton, Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural computation* **22**, 1473–1492 (2010).

66 R Memisevic, Learning to Relate Images. *IEEE Transactions on Pattern Analysis Mach. Intell.* **35**, 1829–1846 (2013).

67 C Finn, I Goodfellow, S Levine, Unsupervised learning for physical interaction through video prediction in *Advances in neural information processing systems.* pp. 64–72 (2016).

68 R Mottaghi, M Rastegari, A Gupta, A Farhadi, "what happens if..." learning to predict the effect of forces in images in *European Conference on Computer Vision.* (Springer), pp. 269–285 (2016).

69 N Watters, et al., Visual interaction networks: Learning a physics simulator from video in *Advances in neural information processing systems.* pp. 4539–4547 (2017).

70 A Nair, et al., Combining self-supervised learning and imitation for vision-based rope manipulation in *2017 IEEE International Conference on Robotics and Automation (ICRA).* (IEEE), pp. 2146–2153 (2017).

71 K Kotovsky, HA Simon, What makes some problems really hard: Explorations in the problem space of difficulty. *Cogn. psychology* **22**, 143–183 (1990).

72 R Primi, Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence* **30**, 41–70 (2001).

73 J Wagemans, et al., A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychol. bulletin* **138**, 1172 (2012).

74 G Kanizsa, *Organization in vision: Essays on Gestalt perception.* (Praeger Publishers), (1979).

75 A Desolneux, L Moisan, JM Morel, *From gestalt theory to image analysis: a probabilistic approach.* (Springer Science & Business Media) Vol. 34, (2007).

76 CB Schönlieb, *Partial differential equation methods for image inpainting.* (Cambridge University Press), (2015).

77 MH Herzog, UA Ernst, A Etzold, CW Eurich, Local interactions in neural networks explain global effects in gestalt processing and masking. *Neural Comput.* **15**, 2091–2113 (2003).

78 C Prodöhl, RP Würtz, C Von Der Malsburg, Learning the gestalt rule of collinearity from object motion. *Neural Comput.* **15**, 1865–1896 (2003).

79 A Amanatiadis, VG Kaburlasos, EB Kosmatopoulos, Understanding deep convolutional networks through gestalt theory in *2018 IEEE International Conference on Imaging Systems and Techniques (IST).* (IEEE), pp. 1–6 (2018).

80 G Ehrensperger, S Stabinger, AR Sánchez, Evaluating cnns on the gestalt principle of closure (2019).

81 B Kim, E Reif, M Wattenberg, S Bengio, Do neural networks show gestalt phenomena? an exploration of the law of closure (2019).

82 DF Bjorklund, *Children's strategies: Contemporary views of cognitive development.* (Psychology Press), (2013).

83 R Siegler, EA Jenkins, *How children discover new strategies.* (Psychology Press), (2014).

84 S Gulwani, et al., Inductive programming meets the real world. *Commun. ACM* **58**, 90–99 (2015).

85 J Hernández-Orallo, SH Muggleton, U Schmid, B Zorn, Approaches and applications of inductive programming (dagstuhl seminar 15442) in *Dagstuhl Reports.* (Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik), Vol. 5.10, (2016).

86 J Hofmann, E Kitzelmann, U Schmid, Applying inductive program synthesis to induction of number series a case study with igor2 in *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz).* (Springer), pp. 25–36 (2014).

87 BM Lake, R Salakhutdinov, JB Tenenbaum, Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).

88 K Ellis, D Ritchie, A Solar-Lezama, J Tenenbaum, Learning to infer graphics programs from hand-drawn images in *Advances in neural information processing systems.* pp. 6059–6068 (2018).

89 D Borrajo, A Roubíčková, I Serina, Progress in case-based planning. *ACM Comput. Surv. (CSUR)* **47**, 35 (2015).

90 JE Laird, et al., Interactive task learning. *IEEE Intell. Syst.* **32**, 6–21 (2017).

91 TR Hinrichs, KD Forbus, X goes first: Teaching simple games through multimodal interaction. *Adv. Cogn. Syst.* **3**, 31–46 (2014).

92 J Kirk, A Mininger, J Laird, Learning task goals interactively with visual demonstrations. *Biol. Inspired Cogn. Archit.* **18**, 1–8 (2016).

93 M Kunda, Nonverbal task learning in *Proceedings of the 7th Annual Conference on Advances in Cognitive Systems (ACS).* (2019).

94 LS DeThorne, BA Schaefer, A guide to child nonverbal iq measures. *Am. J. Speech-Language Pathol.* **13**, 275–290 (2004).

95 M Lázaro-Gredilla, D Lin, JS Guntupalli, D George, Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs. *Sci. Robotics* **4**, eaav3150 (2019).

96 F Chollet, On the measure of intelligence (2019).