Visual Mental Imagery: A View from Artificial Intelligence

Maithilee Kunda (mkunda@vanderbilt.edu)

Department of Electrical Engineering and Computer Science
Vanderbilt University
PMB 351679, 2301 Vanderbilt Place, Nashville, TN 37235-1679

Abstract

This article investigates whether, and how, an artificial intelligence (AI) system can be said to use visual, imagery-based representations in a way that is analogous to the use of visual mental imagery by people. In particular, this article aims to answer two fundamental questions about imagery-based AI systems. First, what might visual imagery look like in an AI system, in terms of the internal representations used by the system to store and reason about knowledge? Second, what kinds of intelligent tasks would an imagery-based AI system be able to accomplish? The first question is answered by providing a working definition of what constitutes an imagery-based knowledge representation, and the second question is answered through a literature survey of imagery-based AI systems that have been developed over the past several decades of AI research, spanning task domains of: 1) template-based visual search; 2) spatial and diagrammatic reasoning; 3) geometric analogies and matrix reasoning; 4) naive physics; and 5) commonsense reasoning for question answering. This article concludes by discussing three important open research questions in the study of visual-imagery-based AI systems—on evaluating system performance, learning imagery operators, and representing abstract concepts—and their implications for understanding human visual mental imagery.

Keywords: knowledge representation; analogical representations; depictive vs. descriptive; iconic vs. propositional; modal vs. amodal.

Kunda, M. (2018). "Visual mental imagery: A view from artificial intelligence." ***Cortex*****, 105, 155-172. https://doi.org/10.1016/j.cortex.2018.01.022**

Contents

1 Introduction **1**

2 A Definition of Visual-Imagery-Based AI **6**

2.1 Three criteria for visual-imagery-based representations 8

2.2 Visual transformations 12

3 A Survey of Visual-Imagery-Based AI Systems **14**

3.1 Template-based visual search 15

3.2 Spatial and diagrammatic reasoning 16

3.3 Geometric analogies and matrix reasoning 20

3.4 Naive physics 23

3.5 Commonsense reasoning for question answering 24

4 Looking Ahead **29**

1 Introduction

> *"I've seen things you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I watched C-beams glitter in the dark near the Tannhäuser Gate. All those moments will be lost in time, like tears in rain."*
>
> – Roy Batty, a replicant
>
> *Blade Runner*

What is the inner, "mental" life of an artificial intelligence (AI) system? At its most basic level, it is true that information in a digital computer is just ones and zeros, but that is a bit like saying that information in the human mind is all just spiking neurons. Humans employ a rich variety of mental representations, ranging from sensory impressions to linguistic symbols, that each can be studied at many different levels of abstraction, e.g., as in Marr's levels of analysis (Marr 1982). And, while some general, low-level principles of operation are shared across different neurons, there is also extensive biological and developmental specialization within the integrated brain-body system that produces very different types of mental representations for different tasks, situations, and sensory modalities.

This article investigates whether, and how, an AI system can be said to use *visual, imagery-based knowledge representations* in a way that is analogous to the use of *visual mental imagery* by people—i.e., using visual, image-like representations to store knowledge, and image-based operations like translation, rotation, and composition to reason about that knowledge in some useful way.

While the existence of visual mental imagery in human cognition was vigorously debated for much of the late 20th century (aptly named "The Imagery Debate"), many convergent findings in neuroscience now support the idea that visual mental imagery is a genuine and useful form of mental representation in humans (Pearson and Kosslyn 2015). Visual mental images are represented in many of the same retinotopic brain regions that are responsible for visual perception, with the key difference that mental images involve neural activations that are not directly tied to concurrent perceptual inputs (Kosslyn, Thompson, et al. 1995; Slotnick, Thompson, and Kosslyn 2005). In addition, the neural activity associated with visual mental imagery has been found to play a functional role: if this neural activity is

artificially suppressed, then a person's performance on certain tasks will decrease (Kosslyn, Pascual-Leone, et al. 1999).

A person's use of visual mental imagery is also associated with certain behavioral characteristics whose study formed much of the early seminal work on this topic in psychology. For example, performing mental rotations of an arbitrary image takes an amount of time that is proportional to the angle through which the rotation is applied, as demonstrated by studies of the now-classic mental rotation task (Shepard and Metzler 1971).

In addition, numerous narrative, often introspective accounts of human intelligence have identified visual mental imagery as playing a crucial role in many different task domains, including medical surgery (Luursema, Verwey, and Burie 2012), mathematics (Giaquinto 2007), engineering design (Ferguson 1994), computer programming (Petre and Blackwell 1999), creativity (Miller 2012), and scientific discovery (Nersessian 2008). Temple Grandin, a professor of animal science who also happens to be on the autism spectrum, identifies her tendency to "think in pictures" as a contributor both to her strengths as a designer of complex equipment for the livestock industry as well as to her weaknesses in understanding abstract concepts and communicating with other people (Grandin 2008). Individuals seem to vary in their abilities to use visual mental imagery from the strong abilities often observed in autism (Kunda and Goel 2011) to the apparent lack of imagery ability recently characterized as *aphantasia* (Zeman, Dewar, and Della Sala 2015).

However, despite the breadth of studies from neuroscience, psychology, and other disciplines, much is still unknown about the cognitive machinery that drives visual mental imagery in humans, such as how mental images are stored in and retrieved from long term memory, how they are manipulated, and how they support intelligent behavior in various real-world task domains. As with research on other aspects of cognition, the study of visual mental imagery is challenging because mental representations and the cognitive processes that use them are not directly observable. We can use neuroimaging to study what happens in the brain, and we can measure behavior to study what happens externally, but the nature of the mental representations themselves can only be inferred indirectly, through these other approaches.

In contrast, the knowledge representations used by an AI system are completely observ-

able. One has only to look up the system's code and inputs, and inspect the state of the system during its operation, to know exactly what knowledge is represented where, and how each piece of knowledge is being used at every moment. For this reason, AI systems are excellent vehicles for conducting scientific, empirical investigations into the relationships between knowledge representations, including the reasoning processes that use them, and intelligent behavior.

In their 1976 Turing Award lecture, AI pioneers Newell and Simon observed that, while computers do play a valuable role as applied tools in people's lives, they also play a valuable role for science and society as objects of empirical inquiry—things that we design, build, and study in order to learn something fundamental about the universe that we live in (Newell and Simon 1976, p. 114):

> Each new program that is built is an experiment. It poses a question to nature, and its behavior offers clues to an answer. Neither machines nor programs are black boxes; they are artifacts that have been designed, both hardware and software, and we can open them up and look inside. We can relate their structure to their behavior and draw many lessons from a single experiment.

Of course, if we studied computers merely to learn more about computers, then the activity would have only so much appeal, but what computers allow us to do is to make empirical study of the more general phenomenon of *computation*. And, to the extent that we believe human intelligence to be at least partly (if not wholly) computational in nature, what AI systems allow us to do is to make empirical study of the phenomenon of computation in the context of intelligent behavior.

But what, exactly, can the study of knowledge representations and reasoning processes in AI systems tell us about mental representations and cognitive processes in people? While some AI systems are designed to realistically model certain human cognitive or neural processes, not all of them are (and in fact probably most are not). *All* AI systems, though, can still tell us something about human intelligence, because each and every one is a small experiment that tests a specific theory of knowledge representation—i.e., the extent to which a particular set of knowledge representations and reasoning processes will lead to a particular

set of outcomes.

Thagard (1996) devised a very nice scheme for describing how such computational theories of representation can be evaluated along five different dimensions, with each contributing in its own way to the study of human cognition (reordered and somewhat paraphrased here):

1. **Psychological plausibility** refers to the extent to which a particular computational theory matches up with what we know about human psychology, for instance in terms of component processes (memory, attention, etc.) or resulting behaviors (reaction times, errors, etc.).
2. **Neurological plausibility** refers to the extent to which a particular computational theory matches up with what we know about the human brain, for instance in terms of functional divisions of the brain or connectionist styles of processing.
3. **Practical applicability** refers to the extent to which a particular computational theory supports useful tools that benefit society, for instance in terms of assistive technologies that help people learn or perform complex tasks.
4. **Representational power** refers to the extent to which a particular computational theory is capable of representing certain classes of knowledge and reasoning. To take a simple example, a representational system consisting only of integers can perfectly represent the number 0 but can only imperfectly represent the number π. Evaluating the representational power of a particular theory in essence asks the question, "What is *possible*, under the terms of this theory?"
5. **Computational power** refers to the extent to which a particular computational theory can support various high-level forms of reasoning, such as planning, learning, and decision making, within reasonable computational bounds of memory and time. Evaluating the computational power of a particular theory in essence asks the question, "What is *feasible*, under the terms of this theory?"

The first two dimensions from this list, psychological and neurological plausibility, are perhaps what come most readily to mind when one thinks of using AI systems to study human cognition. Certain classes of AI systems, e.g., computational cognitive models, biologically-inspired cognitive architectures, etc., are generally evaluated along these two dimensions. Many other classes of AI systems, e.g., self-driving cars, intelligent tutors, applied machine

learning systems, etc., are evaluated primarily along the third dimension, for their practical applicability. The last two dimensions, representational and computational power, are sometimes less explicit in discussions of AI research, though implicitly, the questions of *what is possible* and *what is feasible* drive the design and development of all AI systems.

Here, the contributions of AI systems for understanding human visual mental imagery are discussed primarily in light of these last two dimensions, representational and computational power. Certainly, investigating the degree to which such systems exhibit psychological or neurological plausibility, and how such systems can be of practical benefit to society, are also important, but these questions are not addressed here. Another important factor in recent AI progress, especially in considerations of computational feasibility, has been the rapid expansion of hardware capabilities, especially in hardware optimized for performing many parallel computations. While continued hardware developments are likely to be critical in this and many other areas of AI research, these developments are not discussed here.

This article does aim to answer two fundamental questions about visual-imagery-based AI systems. First, what might visual imagery look like in an AI system, in terms of the internal representations used by the system to store and reason about knowledge? Second, what kinds of intelligent tasks would an imagery-based AI system be able to accomplish? The first question is answered by providing a working definition of what constitutes an imagery-based knowledge representation, and the second question is answered through a literature survey of imagery-based AI systems that have been developed over the past several decades of AI research, spanning task domains of: 1) template-based visual search; 2) spatial and diagrammatic reasoning; 3) geometric analogies and matrix reasoning; 4) naive physics; and 5) commonsense reasoning for question answering. This article concludes by discussing three important open research questions in the study of visual-imagery-based AI systems—on evaluating system performance, learning imagery operators, and representing abstract concepts—and their implications for understanding human visual mental imagery.

2 A Definition of Visual-Imagery-Based AI

In humans, we use the term visual perception to refer to how people see visual information coming in from the outside world, and we use the term visual mental imagery to refer to how people think using visual, image-like internal mental representations. Importantly, visual mental imagery can take place using inputs from visual perception, e.g., being asked to look at and mentally manipulate a given image, or using inputs from other modalities, e.g., creating a mental image from reading text, like: "Visualize a fuzzy yellow kitten."

Unfortunately, in AI, terms like visual thinking, visual intelligence, and visual reasoning are often used interchangeably and confusingly to refer to various forms of visual perception, visual-imagery-based reasoning, or other, non-imagery-based forms of reasoning about visual knowledge. Therefore, in order to clearly define the notion of visual-imagery-based AI, we must first distinguish between the format of an AI system's *input* representations and the format of its *internal* representations.

Just as humans can receive perceptual inputs in many different modalities, an AI system may receive input information in any one (or more) of many different formats, including visual images, sounds, word-like symbolic representations, etc.. Given the information contained in these inputs, the AI system may then convert this information (through "perceptual processing") into one or more different formats to store and reason about this information internally, e.g., as visual images, sounds, word-like symbolic representations, etc. **A visual-imagery-based AI system is one that uses visual images to store and reason about knowledge internally, regardless of the format of the inputs to the system.** Figure 1 shows a simple example of this distinction.

While there have been many AI systems designed to process visual inputs, as demonstrated by the field of computer vision, the vast majority of AI systems designed for non-perceptual tasks use internal representations that are *propositional*, and not visual. Propositional representations are representations in which the format of the representation is independent of its content (Nersessian 2008). Examples of many commonly used propositional representations include logic, semantic networks, frames, scripts, production rules, etc. (Winston 1992). Figure 2 shows an illustration of the "pipeline" of intelligence in a

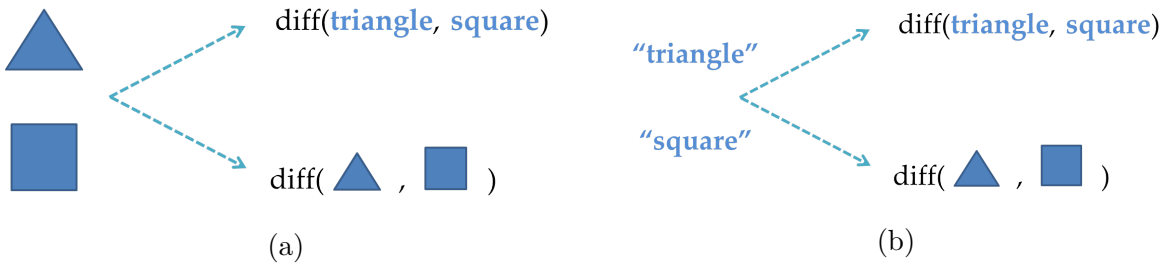


Figure 1: A simple illustration of four different types of AI systems for answering the question, "Are these two shapes the same or different?" (a) The inputs to the AI system are visual images of the two shapes, which can either be converted into internal verbal labels (top) or retained internally as visual images (bottom). (b) The inputs to the AI system are verbal labels of the two shapes, which can either be retained internally as verbal labels (top) or converted into internal visual images (bottom). While all four of these types of systems could be classified as AI systems for visual reasoning, only the two systems illustrated by the bottom pathways would be classified as visual-imagery-based AI systems.

typical propositional AI system. While inputs might initially be received in the form of visual images (or sounds, etc.), they are converted into propositional representations before any reasoning takes place.

In contrast, consider adding a second information pathway to this AI system diagram, as shown in Figure 3. This second pathway illustrates the system's use of visual images as part of its internal knowledge representations. These internal visual images can come from visual inputs (taken as-is or converted into different, perhaps simplified images) or from inputs received in other modalities that undergo conversion into images. Regardless of the input format, reasoning along this pathway can then take place using these internal image representations.

This dual-process pipeline of intelligence allows for the use of both imagistic and propositional representations to solve a given task, very much in the spirit of Paivio's dual-coding theory of mental representations in human cognition (Paivio 2014). Visual-imagery-based AI systems are those that fall into this dual-process category. Some of the AI systems reviewed in this paper use primarily visual-image-based representations, though they might still keep some information (like control knowledge about how to perform a task) represented propositionally. There are also several AI systems that explicitly follow an integrated approach of using both visual and propositional representations of task information, either in sequential

steps or in parallel.

Ultimately, one might expect to see AI systems that use a multi-process approach to intelligence, with access to many different modality-specific pathways of reasoning. In addition, these pathways need not stay separated, as they are shown in Figure 3, but instead can be intertwined, with reasoning mechanisms that can flexibly compare and combine many different types of internal knowledge representations. Such flexibility to move between and combine different kinds of representations is undoubtedly a core aspect of human intelligence, and one that is likely to play an increasingly important role in AI systems in the coming decades.

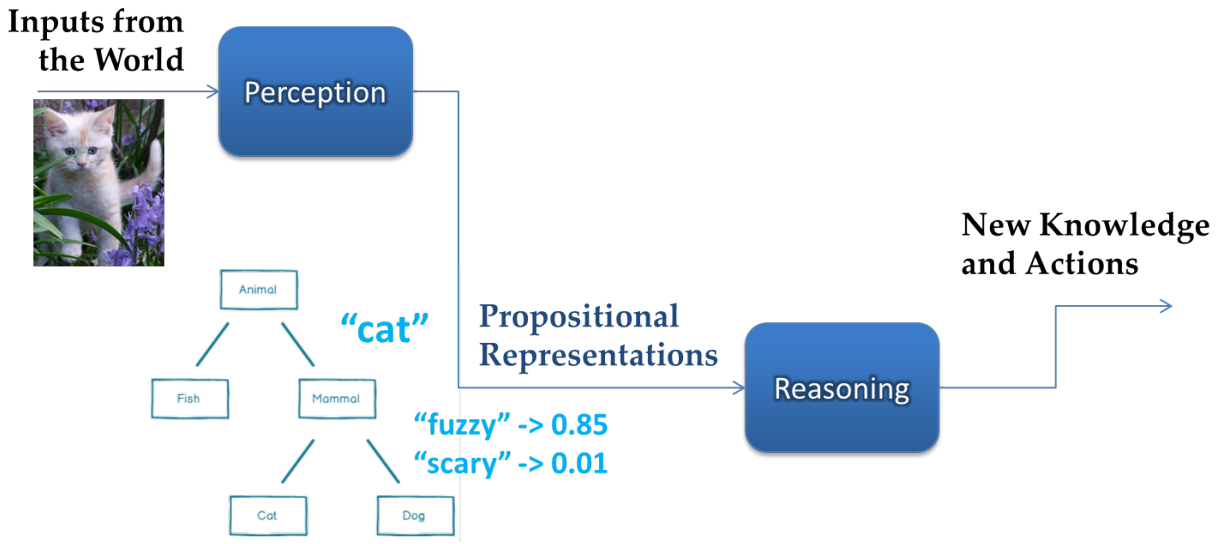


Figure 2: A "propositional pipeline" for intelligent behavior in an AI system. In this simple illustration, the system receives inputs in the form of visual images, which are processed using a perceptual module to extract information that is then stored in various propositional formats. Reasoning takes place over these internal, propositional knowledge representations, in order to produce new knowledge and actions.

2.1 Three criteria for visual-imagery-based representations

In humans, visual mental imagery meets three criteria: 1) the mental representations are image-like, in that they are represented in retinotopically organized brain areas; 2) they do not match concurrent perceptual inputs; and 3) they play some functional role in performing intelligent tasks (Kosslyn, Thompson, et al. 1995; Kosslyn, Pascual-Leone, et al. 1999;

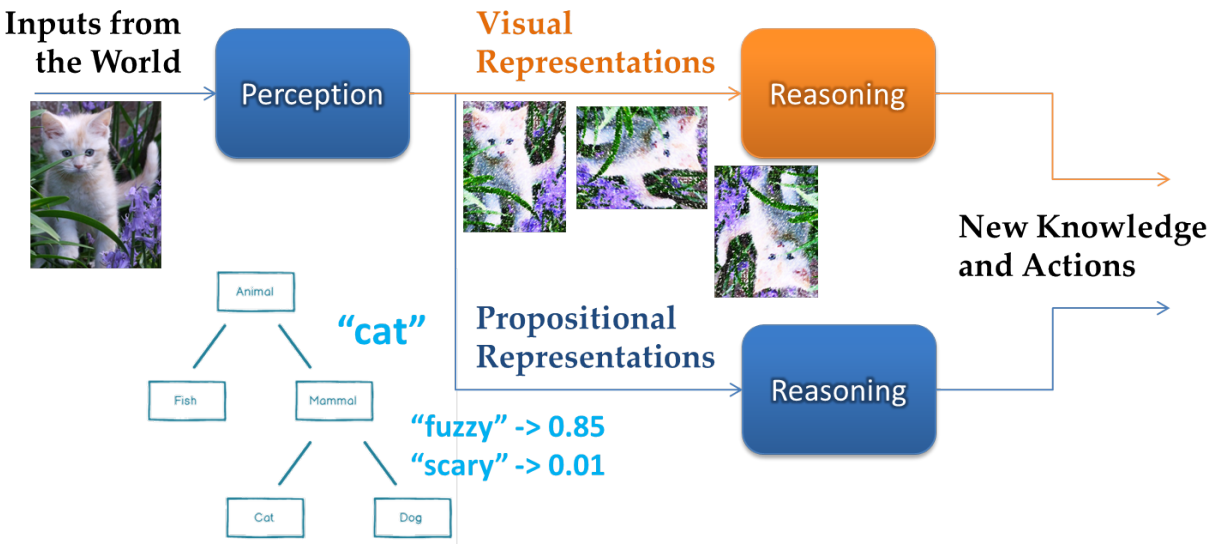


Figure 3: A "dual-process pipeline" for intelligent behavior in an AI system. In this simple illustration, the system receives inputs in the form of visual images, which are processed using a perceptual module to extract information that is then stored and reasoned about either as simplified visual images (top pathway) or in various propositional formats (bottom pathway). Reasoning processes have access to both formats of knowledge representation, in order to produce new knowledge and actions.

Slotnick, Thompson, and Kosslyn 2005). The same three criteria can be adapted to define visual-imagery-based knowledge representations in AI systems.

Criterion 1: Visual-imagery-based representations must be 1) image-like, i.e., iconic, and 2) visual.

While this observation seems simple enough, the question of how to define "image-like" requires some consideration. What makes a knowledge representation image-like is that the representation itself in some way resembles what it represents, i.e., there is some structural correspondence between the format of the representation and its content. Representations that have this property of resemblance or structural correspondence are often called *iconic*, as opposed to propositional representations (as described above) that demonstrate no such correspondence between format and content (Nersessian 2008).

For example, if we consider a picture of a cat, there are spatial relationships in the picture that are the same as the spatial relationships present in the actual cat. The iconic representation does not, of course, preserve every single property of the cat; as all representations are, it is still a simplification and an abstraction (Davis, Shrobe, and Szolovits 1993), but it

is constrained to preserve at least some dimension of information about the cat in a structurally coherent way. The word "cat," on the other hand, is a propositional representation because it preserves no information about the cat in its structure; the relationship between the word and what it represents is completely arbitrary.[1]

The iconic versus propositional distinction often goes by other names.[2] Iconic representations are sometimes called analogical or depictive. Propositional representations are sometimes called descriptive. The iconic property is sometimes defined in terms of homomorphism or isomorphism between the representation and what is represented (Gurr 1998), though many other kinds of definitions have also been proposed (see Shimojima 1999, for a review).

So far, we have defined an imagery-based representation as one that is iconic, but iconic representations do not necessarily have to be visual. In particular, iconic representations can exist in many different modalities, including auditory, haptic, olfactory, etc., and in fact humans do have access to mental imagery in all of these modalities (e.g., Reisberg 2014; Yoo et al. 2003; Stevenson and Case 2005). While these modalities would all be highly interesting to study from an AI perspective, this paper focuses just on imagery in the visual modality, which can be defined as using knowledge representations that are both iconic and that capture appearance-related characteristics (visual and spatial information) of the things that are being represented.

What does this definition look like, in practice? **Iconic visual representations in an AI system are essentially those that are array-based, in which the spatial layout of the array preserves spatial information about what is being represented.**

[1]Linguistic tokens are often, but not always, propositional representations. The linguistic device of onomatopoeia describes one class of words whose phonological structure resembles the auditory properties of their referents. Pictographic or manual alphabets can contain words whose visual structure resembles the visual properties of their referents.

[2]The modal versus amodal distinction is related but refers to a slightly different property of a knowledge representation. A representation is modal if it is instantiated in the same representational substrate that is used during perception (Nersessian 2008). For example, in humans, visual mental imagery would be classified as a modal representation because it is instantiated in many of the same retinotopic brain regions that are used for perception. Amodal representations do not have this property. Classifying representations in an AI system as modal or amodal is not totally straightforward, as what constitutes the system's "perception" is also to some extent a matter of definition. This paper focuses primarily on the iconic versus propositional distinction, with this brief mention of modal versus amodal included mainly as a point of clarification, as the terms have considerable overlap in the literature on knowledge representations.

Individual elements in the array can represent low-level visual features such as intensity, color, or edges—for example, pixel-based RGB images would fall into this category—or individual elements in the array can correspond to higher-level symbolic labels—for example, a simple diagram like `cat-dog-horse` embodies a small set of spatial relationships among the three objects. Such array-based representations can exist in one, two, three, or even four dimensions; an uncompressed movie file is an example of a four-dimensional iconic visual representation.

Criterion 2: Visual-imagery-based representations must differ from concurrent perceptual inputs.

In addition to being iconic and visual, visual-imagery-based representations cannot always match what is coming in through the AI system's "perceptual module," i.e. what is provided to the AI system as input, whether through image sensors or manually fed into the system. This means that the AI system must have some kind of array-based buffer that can store visual information and retain it, even if the visual inputs change or if the inputs are not visual in the first place.

According to this rather basic definition, any AI system that stores any visual images at all would meet this second criterion, and thus could be said to have a rudimentary form of visual-imagery-based representations. To take a slightly more stringent interpretation, we might say that *control* over imagery-based representations cannot come from perception, meaning that the AI system must have some set of internal capabilities for instantiating and manipulating these representations. While the specifics of such capabilities can vary from one system to the next, that these capabilities exist can be considered to be a requirement for visual-imagery-based AI. Examples of commonly implemented capabilities, such as rotation, translation, and composition, are described in Section 2.2 on visual transformations.

Criterion 3: Visual-imagery-based representations must play some functional role in performing intelligent tasks.

Finally, the third criterion requires that the imagery-based representations serve some functional role in intelligent behavior. In an AI system, this means that the representations must contribute in some nontrivial way to solving the task that the system is designed to

address. In humans, some of the most convincing evidence that visual mental imagery serves a functional role, and is not just a byproduct of other reasoning processes, comes from studies that interfere with a person's mental imagery ability using transcranial magnetic stimulation, or TMS (Kosslyn, Pascual-Leone, et al. 1999).

For an AI system, a simple thought experiment that gets at the same issue is to ask, "If we delete the imagery-based representations from this system, would its performance suffer?" This heuristic is especially useful for thinking about many AI systems that claim to model imagery-like processes but use a core set of propositional representations to drive their functionality; these systems often have a "drawing" subroutine that is used only to visualize the reasoning steps to the user, but the images themselves are not actually used for reasoning. Such systems, even though they might be capable of *producing* image-like representations, are not actually *using* these representations to solve the task, and so should not qualify as being imagery-based AI systems.

2.2 Visual transformations

In order to effectively use visual-imagery-based representations to solve a task, an AI system must have not only the ability to create and maintain such representations, but also some means of reasoning about the information contained inside them. In general, systems of knowledge representation are not well specified without the inclusion of a set of valid inference mechanisms that can operate over the symbols in that representation (Davis, Shrobe, and Szolovits 1993). For example, a knowledge representation based on logic should include both the specification of logical symbols as well as rules for deduction.

Studies of visual mental imagery in humans have identified several key inference mechanisms, in the form of visual transformations, that seem to be implicated many different imagery-related task domains:

1. In-plane image translation or scanning (Finke and Pinker 1982; Kosslyn, Ball, and Reiser 1978; Kosslyn 1973; Larsen and Bundesen 1998).
2. Image scaling or zooming, which corresponds to out-of-plane translation (Bundesen and Larsen 1975; Larsen, McIlhagga, and Bundesen 1999).
3. Image rotation (Cooper and Shepard 1973; Zacks 2008).

4. Image composition including intersection (Soulières, Zeffiro, et al. 2011), union (Brandimonte, Hitch, and Bishop 1992b; Finke, Pinker, and Farah 1989), and subtraction (Brandimonte, Hitch, and Bishop 1992b; Brandimonte, Hitch, and Bishop 1992a).

Most of the visual-imagery-based AI systems described in this paper implement some or all of these transformations, though the inclusion of particular transformations and the details of their operation often differ from one AI system to the next. Just as within the world of logic-based representations, there are many different frameworks that have different rules for representation and inference, we need not commit to a single formulation for all imagery-based representations but instead can entertain a variety of different approaches that collectively fall within the category of visual imagery.

As a final comment on transformations, one term often conflated with the use of visual transformations in imagery-based representations is that of *transformation invariance*, which is often discussed in the context of representations used for visual classification. Transformation invariance refers to the ability of a classifier to correctly classify inputs that have been transformed in ways that should not affect the class label. For example, a cat classifier that demonstrates rotation invariance should correctly recognize cats that are upside down, in addition to those that are right-side up. Other commonly discussed types of transformation invariance in visual classification include translation invariance, scale invariance, lighting invariance, etc.

Note that transformation invariance can be achieved using different mechanisms. For example, in order to successfully classify an upside-down cat, a classifier might first apply a rotation to the upside-down cat, and then use an upright-only cat classifier on it. Alternatively, the classifier might have a representation of cats that is intrinsically invariant to rotations, for example by representing cats according to the shapes of their ears, tails, and whiskers, regardless of the orientation of these elements in the image. The latter approach of creating "transformation-invariant representations," i.e., designing the representation itself to be immune to transformations, is a common approach in AI (Kazhdan, Funkhouser, and Rusinkiewicz 2003; Földiák 1991), and aligns with findings from cognitive science that transformation-invariant properties exist in human mental representations (Booth and Rolls 1998).

However, in humans, the two processes of 1) actively applying transformations to mental representations and 2) creating and using representations that have intrinsic transformation-invariant properties, are dissociable (Farah and Hammond 1988) and show distinct patterns of neural activation (Vanrie, Béatse, et al. 2002). Both processes likely play a significant role in the robust visual classification performance than humans are capable of (Tarr and Pinker 1989; Vanrie, Willems, and Wagemans 2001).

Likewise, continued AI research both on applying visual transformations and on creating transformation-invariant representations will likely be valuable in understanding many aspects of visual intelligence. This article focuses primarily on discussions of visual transformations and not of transformation-invariant representations, as the former are more directly relevant to imagery-based representations and reasoning.

3 A Survey of Visual-Imagery-Based AI Systems

This section presents a survey of AI systems that use visual-imagery-based representations, organized by task domain: template-based visual search (Section 3.1), spatial and diagrammatic reasoning (Section 3.2), geometric analogies and matrix reasoning (Section 3.3), naive physics (Section 3.4), and commonsense reasoning (Section 3.5). Each section first describes a few examples of propositional AI approaches that have been developed to solve the given task, and then identifies AI systems that solve these tasks using an imagery-based approach.

Imagery-based AI systems were located by searching the literature using Google Scholar, and especially following reference trails backwards from the later papers as well as forwards from the earlier papers, using Google Scholar's "cited by" function. Where multiple papers appear describing related work from a single research group, one representative paper has been selected for inclusion in this survey. The grouping of AI systems into task domains was done post hoc. While every effort was made to include all published visual-imagery-based AI systems, undoubtedly many have been left out; this survey at least gives a sampling of the AI research that has emerged in this area over the past several decades.

3.1 Template-based visual search

Perhaps the simplest occurrence of visual imagery in AI systems is the use of image templates for visual search. In visual search, a search target must be visually located within a search environment. A very simple visual search task might be to find an instance of the letter "x" somewhere on this page. A more complex visual search task might be to find something in your office to use as an umbrella when it's raining (and when, inevitably, you've left your actual umbrella at home).

During the process of visual search, an AI system can represent the search target in many different ways. In feature-based search, the target is represented by one or more visual features, e.g., "Find the object that is blue and round."

In contrast to feature-based search, an AI system can instead represent the search target using an image that captures aspects of the target's visual appearance. This image is called a *template*, and the corresponding search process is called *template-based search*. A template meets the criteria for being a visual-imagery-based representation, as described in Section 2.1, because it is an iconic visual representation of the search target, it differs from the visual "perceptual" inputs received by the AI system as it inspects the search environment, and it plays a functional role in task performance.

A very simple template-based visual search algorithm might work as follows:

1. Take two images A and B as input, where image A (the template) represents the search target, and image B represents the search environment.
2. Slide the template image A across all possible positions relative to image B. At each position, compute a measure of visual similarity between A and B, for example by calculating a pixel-wise correlation between the two images.
3. Choose the position in image B that yields the highest similarity value to be the final output of the search process.

While this simple algorithm is not particularly efficient or robust to noise, the basic idea of template-based search has been used in many successful AI applications, including recognition of faces (Brunelli and Poggio 1993), traffic signs (Gavrila 1998), medical images (Hill et al. 1994), and more. Extensions to the basic algorithm include more efficient ways

to traverse the search space, such as through the use of gaze or attention models (Rao et al. 2002; Zelinsky 2008; Kunda and Ting 2016; Palmer and Kunda 2018) as well as more flexible ways to represent the template and compute similarity, such as through the use of deformable templates (Yuille, Hallinan, and Cohen 1992).

A complete review of the literature on template-based visual search would be far too long to fit into this paper, and so readers are referred instead to existing reviews (Jain, Zhong, and Dubuisson-Jolly 1998; Brunelli 2009).

3.2 Spatial and diagrammatic reasoning

It might seem like an obvious idea to use visual-imagery-based AI systems for spatial and diagrammatic reasoning tasks. However, the majority of AI systems designed to solve such tasks rely mainly on propositional knowledge representations. (As discussed in Section 2, the vocabulary used by different research groups can be confusing; some groups refer to a "visuospatial reasoning system" to mean an AI system that reasons about visual inputs, regardless of its internal format of representation, while others use the same term to mean a system that reasons using internal visual representations, regardless of the format of the input task. Both might qualify as spatial or diagrammatic reasoning systems, but only the latter would qualify as visual-imagery-based AI under the terms of the criteria outlined in Section 2.1.)

There have been many successful schemes devised for representing visuospatial knowledge in propositional form, for instance by propositionally encoding relations like `is-left-of(X, Y)`. Given such a knowledge representation scheme, an AI system can draw upon this knowledge to make even very complex inferences about a spatial or diagrammatic input problem. For example, one very early effort proposed an AI system that used propositional representations of visuospatial information to generate geometry proofs (Gelernter 1959).

In another early effort, Baylor (1972) built an AI system that reasoned about spatial reasoning problems from a standardized block visualization test. An example problem from this test goes something like this: "Two sides of a 2 inch cube that are next to each other are painted red, and the remaining faces are painted green. The block is then cut into eight 1 inch cubes. How many cubes have three unpainted faces?" Baylor's AI system worked by

first constructing an internal representation of the original block, then performing a "mental simulation" to cut it, and finally inspecting the results to provide the final answer. However, the internal block representations stored by this AI system were *not* iconic; they were stored and accessed as structured lists of vertices, and not as array-based representations. So while this AI system was developed to explore certain problem-solving aspects of "visual mental imagery," its representations were not actually imagery-based in a strict sense.

Continuing in this vein, there have been many successful and interesting propositional approaches to spatial and diagrammatic reasoning demonstrated in AI research. Examples include AI systems that perform qualitative spatial reasoning (Cohn et al. 1997), understand general diagrams (Anderson and McCartney 2003), solve visual analogy problems (Croft and Thagard 2002; Davies, Goel, and Yaner 2008), understand engineering drawings (Yaner and Goel 2008), reason about human-drawn sketches (Forbus et al. 2011), perform path planning (Goel et al. 1994), and many, many more (see Glasgow, Narayanan, and Chandrasekaran 1995 for a review of many of the basic research thrusts in this area). Some approaches to diagrammatic reasoning use graph-based knowledge representations (e.g., Larkin and Simon 1987); while graph-based representations have a bit more internal structure than purely propositional representations, they still do not strictly meet our criteria for imagery-based representations from Section 2.1, as they are not array-based, though it could perhaps be argued that they embody a variant of visual imagery.

There have been far fewer AI systems that perform visuospatial or diagrammatic reasoning using strictly visual-imagery-based representations. The common themes shared by these systems are the use of array-based representations to store iconic visual representations, and the application of visual transformations (e.g., translation, rotation, scaling, etc.) to these array-based representations in order to solve problems from one or more task domains.

Kosslyn and Shwartz (1977) describe an AI system that can construct, inspect, and transform simple images that are stored as unit activations in a 2D matrix, as shown in Figure 4a. Visual transformations include translation, scaling, and rotation. This system does not solve any particular task, per se, but was developed to elucidate some basic computational processes of visual imagery.

Mel (1990) describes an imagery-based AI system used in motion planning for a robot

arm, in which the robot first learns mappings between its commanded servo outputs and its own visual percepts of the movements of its arm, and then plans new motions essentially by generating and inspecting new internal images of how it wants its arm to move.

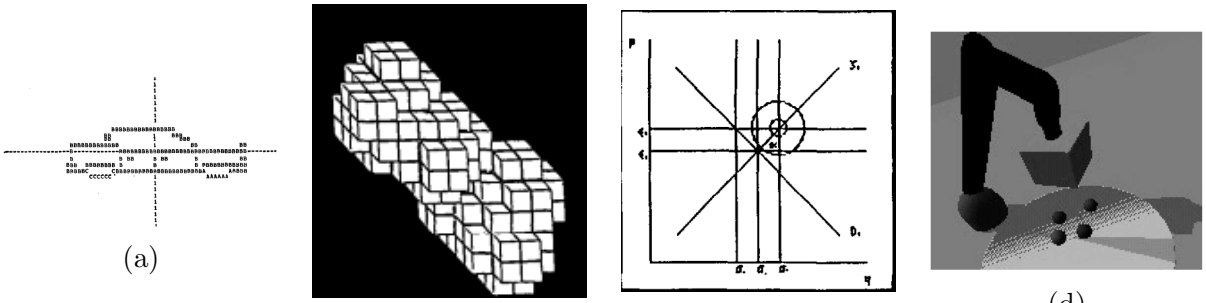
Glasgow and Papadias (1992) present one of the better known works on imagery-based AI. They describe a system that uses nested arrays to store imagery-based representations at multiple levels of abstraction. At the lowest level, 3D arrays serve as iconic representations of shape and are used for problem solving in task domains like 3D molecular shape analysis, as shown in Figure 4b.

Tabachneck-Schijf and colleagues (1997) describe an AI system called Computation with Multiple Representations (CaMeRa) that uses both propositional and imagery-based representations to interpret 2D line graphs in the domain of economics. The CaMeRa system has a visual buffer that uses array-based representations and transformations to "visually" trace different imagined lines on a graph. For instance, in order to detect where some point lies relative to the x-axis of the graph, the system essentially visualizes a vertical line coming down from the point and then observes where this line crosses the x-axis, all within its visual buffer. Figure 4c shows an illustration of the visual buffer in the CaMeRa system.

Roy and colleagues (2004) describe an imagery-based module for a robotic arm that enables the robot to reason about differing visual perspectives of its own environment. As shown in Figure 4d, the robot generates a visual image that depicts the scene in front of it (objects on a table) from the perspective of a human sitting across the table; in this view, the robot is visualizing not just how the objects look to the human but also its own appearance.

Lathrop and colleagues (2011) implemented a visual imagery extension to the well known SOAR cognitive architecture. The resulting system uses imagery-based representations to solve problems in a simple block-stacking task domain as well as in a more complex, multi-agent mapping and scouting task domain. In both domains, the system visualizes the results of its actions before it executes them, in order to help in planning and action selection.

Other AI systems for spatial and diagrammatic reasoning that include some visual-imagery-based representational component include NEVILLE by Bertel and colleagues (2006), DRS by Chandrasekaran and colleagues (2011), PRISM by Ragni and Knauff (2013), and Casimir by Schultheis and colleagues (2011; 2014).

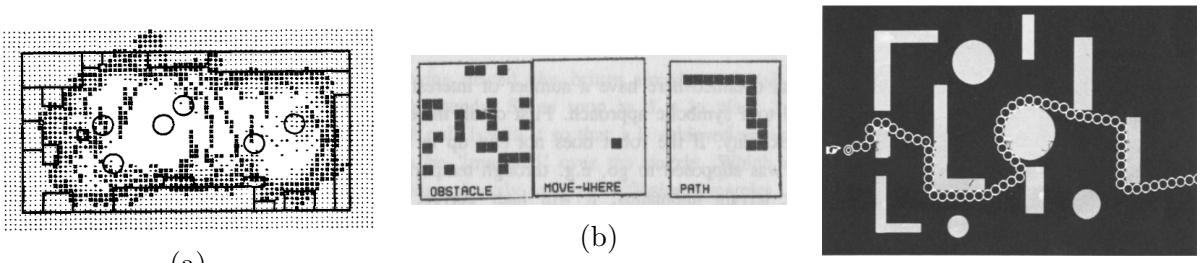


(a) (b) (c) (d)

Figure 4: Examples of internal, visual-imagery-based representations used by AI systems for spatial or diagrammatic reasoning tasks. (a) Kosslyn and Shwartz 1977. (b) Glasgow and Papadias 1992. (c) Tabachneck-Schijf, Leonardo, and Simon 1997. (d) Roy, Hsiao, and Mavridis 2004.

One kind of spatial reasoning task worth noting separately is that of reasoning about maps. There are many ways for an AI system to store map-like information, including as a set of propositionally represented statements (e.g., Myers and Konolige 1994). Occupancy grids, now a very common approach, were first introduced by Moravec and Elfes (1985) as a way for mobile robots to aggregate and store information about a new environment during exploration, as shown in Figure 5a. An occupancy grid is a 2D or 3D array-based data structure that corresponds to a map of the environment; the contents of each cell reflect the robot's estimate of what exists at the corresponding location in the actual environment. Many approaches in robotics, such as Kuipers' (2000) Spatial Semantic Hierarchy, combine occupancy-grid-based and propositional map representations.

Occupancy grids meet the requirements for an imagery-based representation because they are iconic and often visual (though some occupancy grids may capture non-visual information about the environment as well), they do not correspond directly to any single visual percept received by the robot, and they play a functional role in the robot's spatial reasoning. In many occupancy-grid-based approaches, while the grid itself might be stored in an imagery-based way, the inference operations performed over these representations (like planning a shortest path between two points) are often defined in terms of graph algorithms and not in terms of visual transformations. However, there have been at least two attempts to devise path planning algorithms that use visual transformations over occupancy grids, as shown in Figures 5b and 5c (Steels 1988; Gardin and Meltzer 1989).



(a) (b) (c)

Figure 5: Examples of internal imagistic representations used by AI systems for mapping and path planning. (a) Moravec and Elfes 1985. (b) Steels 1988. (c) Gardin and Meltzer 1989.

3.3 Geometric analogies and matrix reasoning

Geometric analogies are a class of problems often found on human intelligence tests that follow the standard analogy problem format of, "A is to B as C is to what?" In a geometric analogy problem, A, B, and C are all images, and the correct answer must be selected from a set of possible choices, as shown on the left of Figure 6. Matrix reasoning problems are similar; a matrix of images is presented with one missing, and the correct missing image must be selected from a set of possible choices, as shown on the right of Figure 6.

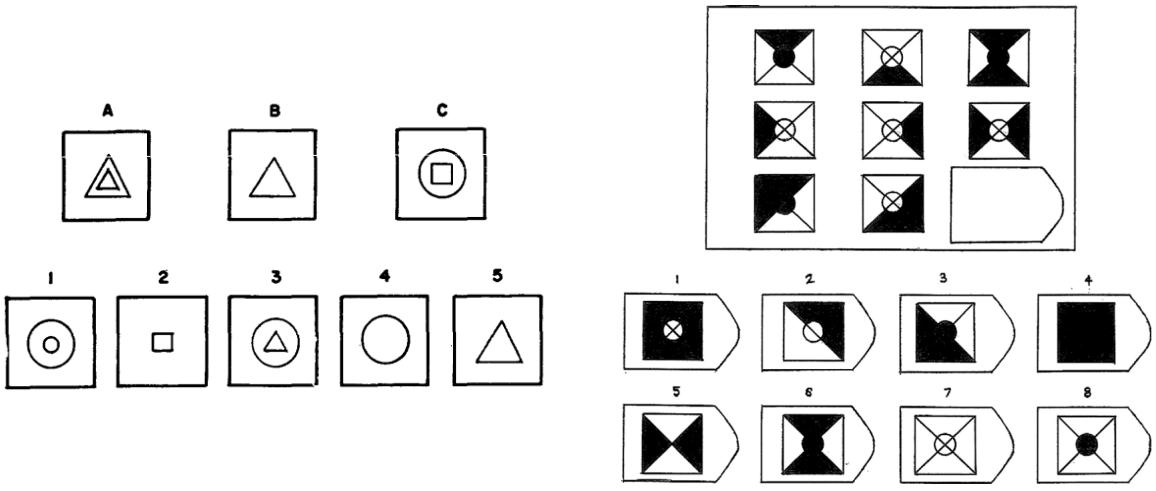


Figure 6: Left: Example geometric analogy problem (Evans 1968). Right: Example matrix reasoning problem similar to those found on the Raven's Progressive Matrices tests (Kunda, McGreggor, and Goel 2013).

Both of these types of problems have appeared on human intelligence tests for decades. One such series of matrix reasoning tests, the Raven's Progressive Matrices, are used as

standardized measures of fluid intelligence in numerous clinical, scientific, and educational settings (Raven, Raven, and Court 1998), and in fact the Raven's tests have been identified in the field of psychometrics as being the best single-format measure of general intelligence that exists (Snow, Kyllonen, and Marshalek 1984).

Evans (1968) demonstrated an AI system called ANALOGY that solves geometric analogy problems using propositional representations. ANALOGY contains a perceptual module that takes line descriptions of a geometric analogy problem as input and produces propositional list-based representations of the problem as output, which are then used by ANALOGY during the rest of the solution process. For example, the first image A in the geometric analogy problem shown on the left of Figure 6 might be converted into something like:

```
((P1 P2) (INSIDE P2 P1) (P1 P2 ((1.2) .  (0.0) .  (N.N.))))
```

This representation roughly translates to saying, "There are two figures, P1 and P2. P2 is inside P1. P1 is 1.2 times larger than P2, the relative rotation between P1 and P2 is 0.0 degrees, and there are no reflection relationships between P1 and P2."

Many subsequent AI systems have used similar formats of propositional representations to solve both geometric analogy and matrix reasoning problems, investigating many interesting aspects of this task domain including maintaining goals and subgoals in working memory (Carpenter, Just, and Shell 1990), logical reasoning techniques (Bringsjord and Schimanski 2003), techniques for analogical mapping between problem elements (Lovett et al. 2009), representing hierarchical patterns in problem information, (Strannegård, Cirillo, and Ström 2013), and the induction of solution rules (Rasmussen and Eliasmith 2011).

However, these propositional AI systems do not explain a different type of solution strategy that humans can and do use, which is to recruit visual mental imagery instead of relying purely on propositional (e.g., verbal or linguistic) mental representations. There is strong evidence that humans generally use a combination of imagery-based and propositional representations to solve these kinds of problems (DeShon, Chan, and Weissbein 1995; Prabhakaran et al. 1997). (See Kunda, McGreggor, and Goel 2013 for a much more detailed review of the literature on both human and AI problem-solving strategies on the Raven's tests.)

Early theoretical work in AI suggested the kinds of algorithms that might play a role

in imagery-based solution strategies to matrix reasoning problems, though the algorithms were not implemented in an actual system (Hunt 1974). More recently, Kunda and colleagues (2013) constructed an AI system called the Affine-and-Set Transformation Induction (ASTI) system that uses visual images to represent information from matrix reasoning problems, and reasons about these images using imagery operations such as translation, rotation, and composition. The ASTI system meets our criteria for a visual-imagery-based AI system because 1) it uses iconic visual representations of problem information, 2) these images differ from perceptual inputs because they are translated, rotated, and otherwise altered to form new images that are not contained anywhere in the original problem, and 3) the images play a functional role in the system's problem-solving procedures.

To solve a matrix reasoning problem, the ASTI system follows a problem-solving approach called constructive matching (Bethell-Fox, Lohman, and Snow 1984). First, the ASTI system tries out a series of imagery operators on different images from the original problem matrix until it finds an operator that can "visually simulate" the change that occurs across any single row or column of the matrix. Then, it uses this operator to construct a new image that fits in the blank space of the matrix. Finally, it compares this constructed answer to the list of answer choices in order to select the most visually similar answer choice.

The ASTI system was tested against the Standard version of the Raven's Progressive Matrices series of tests, which is of medium difficulty and is intended for children and adults of average ability. Out of 60 total problems on the test, the ASTI system answered 50 correctly, which is around the level of performance expected for typically developing 16-17-year-olds (Kunda 2013). This result was the first concrete proof that it is possible (from a computational perspective) to get a score of 50 using a purely imagery-based approach. Prior to this finding, a common belief about the Raven's tests was that imagery-based reasoning could only solve the very easiest problems, and that solving the harder problems required switching to a propositional strategy (Hunt 1974; Kirby and Lawson 1983). The ASTI result also lends weight to findings that certain individuals on the autism spectrum appear to rely more heavily on visual brain regions when solving Raven's problems than do neurotypical individuals, with no decrease in accuracy (Soulières, Dawson, et al. 2009).

A related, parallel AI effort by McGreggor and colleagues (2014) investigated imagery-

based reasoning on the Raven's test using fractal image representations, which involve using imagery-like operations to construct representations of problem information that capture similarity and self-similarity at multiple spatial scales across different sets of input images. These fractal image representations were used as part of an AI system that solved Raven's test problems (McGreggor and Goel 2014) as well as visual odd-one-out problems (McGreggor and Goel 2011), and the method was also later applied to analogy-based task transfer in robotics (Fitzgerald et al. 2015).

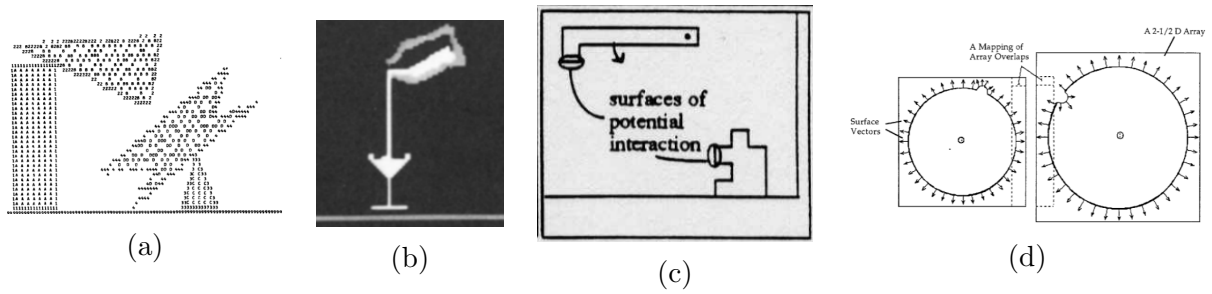
3.4 Naive physics

How do intelligent systems (human or AI) represent and reason about the physical nature of the world? Clearly, one does not need to know the correct Newtonian physics equations in order to predict that a ball will roll down a hill. Early work in AI proposed the use of qualitative representations of physics knowledge to support fast, approximate "naive physics" reasoning. For example, instead of representing the exact volume of liquid in a glass of water, we might think of it as being completely full, mostly full, mostly empty, etc. These approximations are "close enough" to generate successful answers to many questions about what will happen to this glass water in different situations. Many AI systems have adopted such propositional forms of representation to reason about qualitative physics concepts (Forbus 1984; De Kleer and Brown 1984).

While these AI systems were intended primarily as models of human reasoning, other areas of computer science developed techniques of physics-based modeling, i.e., using quantitative propositional representations to simulate physical situations, using physics equations as the core form of knowledge in the computer system. Some recent work blends these two by proposing simulation-based models of naive physics reasoning (Johnston and Williams 2009), including proposals that perhaps humans use some form of simulation-based reasoning as well as qualitative reasoning, though the format of the core physics knowledge in humans is still an open question (Hamrick, Battaglia, and Tenenbaum 2011).

A third view is that naive physics reasoning in humans might be based on internal simulations that are not mathematically defined but rather visually defined, i.e., using visual mental imagery. In line with this view, Funt (1980) presented an AI system called WHISPER

that used interactions between neighbors in a connected network of units to simulate basic physical processes in a block world domain, such as object stability and toppling, as shown in Figure 7a. Gardin and Meltzer (1989) developed an AI system that uses an imagery-based representation formed of connected units that simulates flexible objects like rods of varying stiffness, strings, and liquids by changing parameters on the unit connections, as shown in Figure 7b. Shrager (1990) described an AI system that uses a combination of imagery-based and other representations to reason about problems in a gas laser physics domain. Narayanan and Chandrasekaran (Narayanan and Chandrasekaran 1991) described an AI system that also uses a combination of imagery-based and other representations to reason about blocks-world problems, as shown in Figure 7c. Schwartz (Schwartz and Black 1996) described an AI system that models unit forces in array-based representations in order to simulate the rotations of meshed gears, as shown in Figure 7d.



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

Figure 7: Examples of visual-imagery-based representations used by AI systems for reasoning about naive physics concepts. (a) Funt 1980. (b) Gardin and Meltzer 1989. (c) Narayanan and Chandrasekaran 1991. (d) Schwartz and Black 1996.

3.5 Commonsense reasoning for question answering

In AI, commonsense reasoning capabilities are held to be critical to virtually every area of intelligent behavior, including question answering, story understanding, planning, and more (Davis 2014). However, commonsense reasoning remains a difficult challenge for the field. For example, answering certain questions—e.g., "Could a crocodile run a steeplechase?"—is easy for many people but difficult for most AI systems, requiring not only language processing but also everyday background knowledge that is hard to encode (Levesque 2014). Answering these kinds of "commonsense" questions has been proposed as an alternative

to the Turing test as a way to characterize the extent to which a machine demonstrates intelligence (Levesque, Davis, and Morgenstern 2011).

Over the past few decades, there have been several massive projects undertaken to construct AI systems that perform commonsense reasoning using propositional representations of background knowledge. Much of the effort in these projects has gone into essentially writing down huge amounts of commonsense knowledge in specialized, interconnected, machine-interpretable formats, as well as into developing scalable search and reasoning algorithms that can pull this knowledge together to answer specific questions that are presented to the system.

Lenat's CYC system (short for "encyclopedia"), begun in 1984, recruited teams of people to manually enter knowledge statements into the CYC database. Another system called Open Mind Common Sense was an early adopter of the crowdsourcing philosophy, recruiting volunteers over the Internet to contribute knowledge statements (Singh et al. 2002). More recently, there have been many AI efforts aimed at automatically extracting structured knowledge from existing Internet sources such as Wikipedia (Ponzetto and Strube 2007). IBM's Watson system, while not focused specifically on commonsense reasoning per se, defeated reigning human champions on the game show Jeopardy! by drawing from "a wide range of encyclopedias, dictionaries, thesauri, newswire articles, literary works, and so on" (Ferrucci et al. 2010, p. 69).

All of these approaches use propositional representations of knowledge to process incoming language, reason about the given information, and answer questions about what has been described. However, another way to approach this kind of task could be to create a visual image of the situation and then use visual imagery operators to manipulate and query the image in order to obtain the desired information. For example, in response to the crocodile-steeplechase question, one can visually imagine a crocodile running a steeplechase and then evaluate how reasonable the scene looks by "inspecting" the generated visual mental image. Perlis (2016) emphasizes the importance of building AI systems that incorporate this "envisioning" approach to planning and understanding. Winston conceptualizes this type of reasoning as a capability that combines both imagery and storytelling, often presenting his own table-saw example as a thought experiment (Winston 2012, p. 25):

> As a friend helped me install a table saw, he said, "You should never wear gloves when you use this saw." At first, I was mystified, then it occurred to me that a glove could get caught in the blade. No further explanation was needed because I could imagine what would follow. It did not feel like any sort of formal reasoning. It did not feel like I would have to have the message reinforced before it sank in. It feels like I witnessed a grisly event of a sort no one had ever told me. I learned from a one-shot surrogate experience; I told myself a story about something I had never witnessed, and I will have the common sense to never wear gloves when I operate a table saw.

There have been numerous AI systems developed over the years that aim to answer commonsense-type questions using visual-imagery-based representations. Not surprisingly, early work in this area focused on using imagery-based representations to represent and answer questions specifically about spatial relationships in natural language sentences. In one of the earliest published papers on this topic, Waltz and Boggess (1979) describe an AI system that constructs 3D descriptions of objects and their relationships, and then uses these 3D descriptions to answer questions about the scene. However, this system stores objects internally as sets of numerical coordinates, and the "image" is accessed only implicitly through calculations about these coordinate values, and so the system does not strictly meet the criteria for imagery-based representations laid out in Section 2.1.

Many of the other AI systems described in this section similarly use coordinate-based descriptions of scene models. For example, if a 3D modeling engine is used (as is often the case) to generate scene descriptions, the internal representation used by the AI system is the native representation of the 3D modeling engine, which is often coordinate-based. These systems fall into somewhat of a grey area regarding imagery-based AI; the spirit of the approach is certainly imagery-like, but the internal representations used by these systems do not always strictly meet the criteria for visual-imagery-based representations described in Section 2.1. Regardless, this general area of research is certainly an important one for the continued development of imagery-based AI systems, and so this section includes AI systems that are either strictly imagery-based or at least imagery-based in spirit. Certainly all of these AI systems can produce new images as outputs (Criterion 1), and these images do not

match any perceptual inputs of these systems (Criterion 2), as shown in Figure 8; all that remains is for the system to have some reasoning procedure that operates directly on these images to solve a particular problem (Criterion 3).

A typical AI system in this category is often set up as a question-answering system. The input to the system is a text description of some situation or scene, along with a question about the scene. The system should be able to output the correct answer to the question. This kind of system is often designed to function using three distinct modules:

1. A natural language module converts the input text (both the scene description and the question) into structured, propositional descriptions, for example in the form of logical statements.
2. An imagery module converts the structured descriptions of the scene into a 2D or 3D scene image that depicts the given scene information.
3. Based on the contents of the question, a reasoning module inspects the scene image to obtain whatever information is necessary to answer the question.

The first part of this process falls into the category of natural language processing (NLP), a very broad area of AI. For AI systems that aim to create a visual image from given language, the language processing step is often specifically geared towards extracting spatial and temporal relationships.

The second part of this process, constructing an imagined scene, requires that the system already encodes background knowledge about what different scene objects and relationships mean. Many systems rely on a predefined knowledge database that contains default object models (e.g., a 3D model of a typical table) used to construct the scene. One of the main technical challenges that such systems must solve is how to reconcile the ambiguity present in a textual scene description with the specificity of a concrete scene image; solutions include generating multiple possible scene images (Ioerger 1994) or probability distributions over where objects might be located (Schirra and Stopp 1993). Some systems attempt to address the research question of where this knowledge database comes from, i.e., how this knowledge can be learned from experience (Schirra and Stopp 1993; Chang, Savva, and Manning 2014). Figure 8 shows snapshots from the imagined scenes of several different AI systems that take input language and convert the given information into new 2D or 3D scene images.

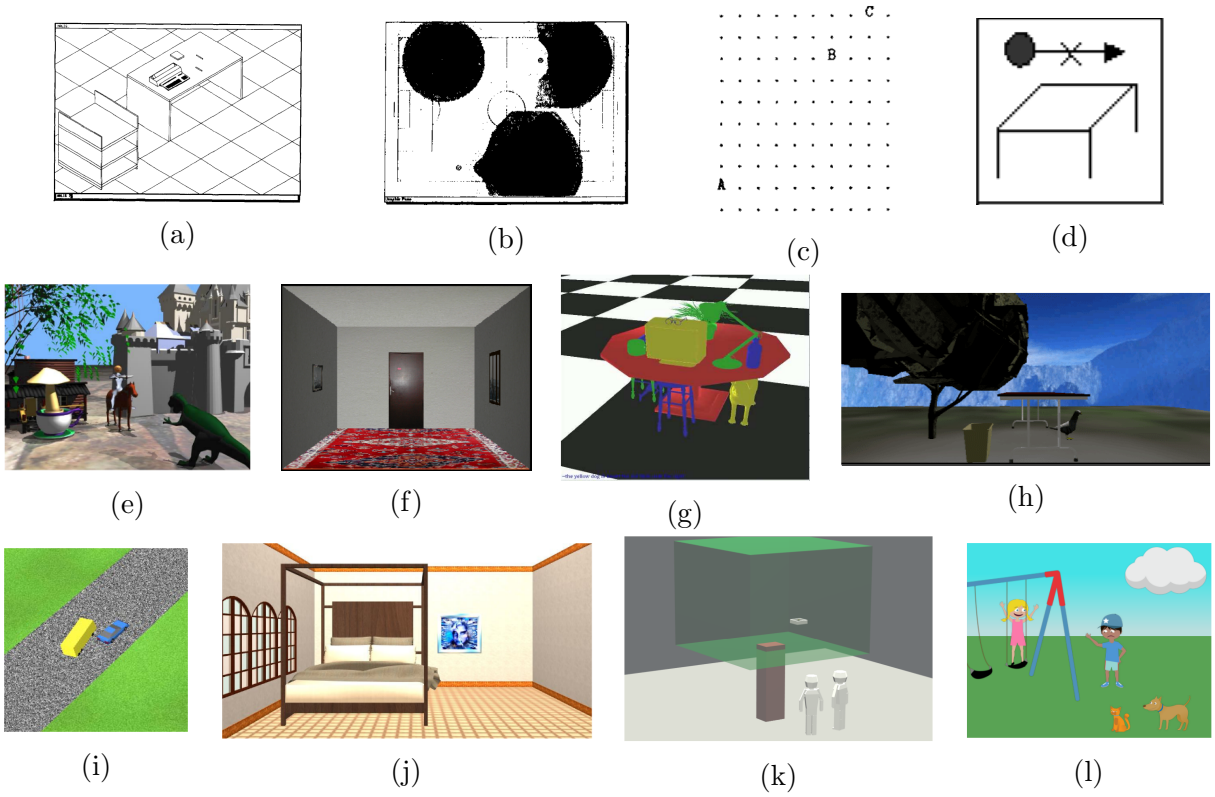


Figure 8: Examples of visual-image-based scenes constructed by AI systems based on text-only inputs. (a) Giunchiglia et al. 1992. (b) Schirra and Stopp 1993. (c) Ioerger 1994. (d) Bender 2001. (e) Coyne and Sproat 2001. (f) Durupınar, Kahramankaptan, and Cicekli 2004. (g) Seversky and Yin 2006. (i) Johansson et al. 2005. (h) Finlayson and Winston 2007. (j) Chang, Savva, and Manning 2014. (k) Bigelow et al. 2015. (l) Lin and Parikh 2015

This kind of scene construction by an AI system is sometimes called "text-to-scene" conversion. In this literature, the generated scenes are sometimes intended to be for human consumption, for instance as automated story illustration systems. Such systems may end with the second part of the process, scene generation, and not perform any subsequent reasoning over the generated image. However, these systems do address many central research questions relevant to general imagery-based AI, such as how visual background knowledge can be encoded, how linguistic ambiguities can be resolved, etc.

The third part of the process involves reasoning about the imagined image, often to answer a question that was received as part of the system's inputs. Here, the concrete nature of the imagined image (which poses such challenges in image creation) is what gives a great advantage for reasoning, because there is much information about the scene that

was not explicitly described in the initial text description but is now available for immediate querying by the reasoning module.

To take a simple example, suppose we have two statements, "The fork is left of the plate," and, "The plate is left of the knife." Is the fork left of the knife? For an AI system storing the initial statements in propositional form, even though the information is sufficient to answer the question, the answer is not immediately available; some type of inference must chain together the two statements in order to compare the two objects. However, for an AI system storing the initial statements as a concrete image, the information about the relative position of the fork and knife, though never explicitly stated in the input text, is available for immediate inspection. While in this simple example, there might not be much computational difference between the two approaches, consider what happens if we are chaining together a dozen object statements, or a hundred, or a million. While propositional representations certainly have other advantages, this particular type of gain in reasoning efficiency for imagery-based representations has been acknowledged in AI (Larkin and Simon 1987).

4 Looking Ahead

While there has been much progress made in visual-imagery-based AI systems over the past several decades, as evidenced by the survey presented in Section 3, there is still much to be learned about the computational underpinnings of visual imagery and their role in intelligence. What follows is a brief discussion of three important open research questions in the study of visual-imagery-based AI systems.

How can imagery-based AI systems be evaluated? For many task domains, it is easy to set up objective tests to evaluate how well an AI system is performing. Natural language understanding can be tested by having conversational interactions with an AI system, or by having it process a piece of text and respond to queries afterwards. Visual perception can be tested by showing the AI system images or videos, and then having it identify what it has seen. How does one test the visual imagery capabilities of an AI system? Most of the imagery-based AI systems discussed in Section 3 were designed to solve problems from a

particular task domain. Some published studies describe quantitative results obtained from testing the AI system against a comprehensive set of such problems; other studies describe only a few results from testing the AI system against representative example problems, and still others present a proof-of-concept of the AI system with little to no testing.

While there has been an impressive breadth of research across different task domains, as evidenced by the survey in Section 3, there has not yet been the kind of decades-long, sustained research focus that has yielded deep AI insights in other areas, such as, for example, in computer vision, which has involved many hundreds of research groups around the world studying closely related problems in visual recognition, segmentation, etc. One issue is that visual mental imagery in humans is itself difficult to study, with no standardized tests of imagery ability in wide use. Also, many imagery-related tasks in people are either too easy (e.g., mental rotation) or too difficult (e.g., imagining a table saw) to readily tackle as an AI research project.

Following the example of computer vision, standardized benchmarks of the right difficulty level can help generate a critical mass of research in a particular task domain, though of course benchmarks present their own set of issues related to evaluation. Whether through benchmarks or perhaps more systematic designs of individual research studies, there is significant need and opportunity for advancing evaluation methods for imagery-based AI systems.

How are imagery operators learned? In humans, the reasoning operators used during visual mental imagery (visual transformations like mental rotation, scaling, etc.) are believed to be learned from visuomotor experience, e.g., watching the movement of physical objects in the real world (Shepard 1984). However, we still have no clear computational explanation for how this type of learning unfolds. Mel (1986) proposed an ingenious method for the supervised learning of visual transformations like rotation from image sequences; in this approach, each transformation operator is represented not as a single image function but instead as a set of weights in a connectionist network, i.e., a representation that is both distributed and continuous. Then, weights in this network are updated according to a standard perceptron update rule. Mel implemented an AI system called VIPS that successfully learned simple operators from simulated wireframe image sequences depicting the given transformations. Memisevic and Hinton (2007; 2010) demonstrate an approach that

uses more complex connectionist networks to learn several different transformations in an unsupervised fashion from large video databases. Seepanomwan and colleagues (2013) propose a robot architecture that successfully combines visual and motor perceptual information to learn mental rotation by rotating objects and watching how their appearance changes in a simulated environment.

While many AI systems implement visual transformations as distinct operations comprising a finite "imagery operator" library (Kunda, McGreggor, and Goel 2013, e.g.), another possibility is that continuous operators could be represented in terms of distinct, infinitesimal basis functions that can be combined in arbitrary ways (Goebel 1990). We still do not know exactly how humans represent the transformations used in visual mental imagery, though there is evidence that operators like mental rotation are sometimes easier for people to perform along primary axes than off-axis (Just and Carpenter 1985). Recent AI advances in deep learning, if applied to the problem of learning imagery operators, may help to identify effective forms of low-level representations that facilitate this particular kind of learning (Bengio, Courville, and Vincent 2013).

The question of how imagery-related reasoning skills are learned is crucial not only for research in AI but also for human education; visuospatial ability is increasingly viewed as a key contributor to math learning (National Research Council 2009; Cheng and Mix 2014) and to success in many STEM fields (Wai, Lubinski, and Benbow 2009). Moreover, recent research suggests that many different visuospatial abilities can be improved with training (Uttal et al. 2013). While it is generally agreed that people learn imagery-based reasoning skills through perceptual experience, it is less clear what types of experience are most valuable, and why, and how to design training interventions that precisely target these learning experiences. AI systems are already used in many different education domains to improve student learning outcomes, and so perhaps imagery-based AI systems could serve as tools for improving math and STEM learning by helping pinpoint how best to boost a person's imagery-related reasoning skills.

How can imagery-based representations be used to reason about abstract concepts? Most of the imagery-based AI systems listed in Section 3 use their imagery-based representations to reason about information that is essentially visual. Even for systems that

have non-visual inputs, such as the commonsense reasoning systems described in Section 3.5, the knowledge that is being represented is generally about things like spatial relationships, the visual appearance of semantic categories, etc. However, in humans, many interesting examples of visual mental imagery involve reasoning about information that is inherently abstract and non-visual. For example, both Albert Einstein and Richard Feynman observed that they often thought about abstract physics concepts first using visual mental images, and only afterwards using equations Gleick 1992; Feist 2008. As Feynman once described to an interviewer Gleick 1992, p. 244:

> 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.

Part of what humans do so marvelously is take cognitive processes that may have originally evolved for one purpose (e.g., using visual mental imagery to reason about space) and use them for something else entirely (e.g., usingvisual mental imagery to reason about abstract mathematical concepts)—a sort of metaphorical thinking (Lakoff and Johnson 2008). Can imagery-based AI systems ever tackle the deep thoughts of scientists like Feynman and Einstein? Polland (1996) compiled an extensive list of mental imagery reports from biographical and autobiographical accounts of 38 famous scientists, artists, musicians, and writers, and analyzed what role mental imagery seemed to play in the creative problem-solving processes of each subject. Perhaps someday, imagery-based AI systems could help to explain the computational mechanisms behind these kinds of advanced, open-ended, and creative problem-solving episodes by some of our greatest thinkers.

Acknowledgment

This research was supported in part by the National Science Foundation (Grant #1730044) and by the Vanderbilt Discovery Grant program (2017 award). Many thanks to Ashok K.

Goel, Keith McGreggor, and David Peebles for helpful discussions and for their comments on this manuscript; to Isabelle Soulières, Michelle Dawson, and Laurent Mottron for sharing their insightful research on visual cognition in autism; and to Temple Grandin for her eloquent writings on "Thinking in Pictures" that initially inspired this research.

References

Anderson, Michael and Robert McCartney (2003). "Diagram processing: Computing with diagrams". In: *Artificial Intelligence* 145.1-2, pp. 181–226.

Baylor, George W (1972). "A treatise on the mind's eye: An empirical investigation of visual mental imagery." PhD thesis. Carnegie Mellon University.

Bender, John R (2001). "Connecting Language and Vision Conceptual Semantics". MA thesis. Massachusetts Institute of Technology.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828.

Bertel, Sven et al. (2006). "Sketching mental images and reasoning with sketches: NEVILLE–a computational model of mental & external spatial problem solving". In: *Proceedings of the 7th International Conference on Cognitive Modeling, Trieste (ICCM 2006)*, pp. 349–350.

Bethell-Fox, Charles E, David F Lohman, and Richard E Snow (1984). "Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance". In: *Intelligence* 8.3, pp. 205–238.

Bigelow, Eric et al. (2015). "On the need for imagistic modeling in story understanding". In: *Biologically Inspired Cognitive Architectures* 11, pp. 22–28.

Booth, MC and Edmund T Rolls (1998). "View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex." In: *Cerebral Cortex* 8.6, pp. 510–523.

Brandimonte, Maria A., G J Hitch, and D V Bishop (1992a). "Influence of short-term memory codes on visual image processing: evidence from image transformation tasks". eng. In:

Journal of experimental psychology. Learning, memory, and cognition 18.1, pp. 157–165. ISSN: 0278-7393.

Brandimonte, Maria A., G J Hitch, and D V Bishop (1992b). "Manipulation of visual mental images in children and adults". eng. In: *Journal of experimental child psychology* 53.3, pp. 300–312. ISSN: 0022-0965.

Bringsjord, Selmer and Bettina Schimanski (2003). "What is artificial intelligence? psychometric AI as an answer". In: *Proceedings of the 18th international joint conference on Artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 887–893.

Brunelli, Roberto (2009). *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons.

Brunelli, Roberto and Tomaso Poggio (1993). "Face recognition: Features versus templates". In: *IEEE transactions on pattern analysis and machine intelligence* 15.10, pp. 1042–1052.

Bundesen, Claus and Axel Larsen (1975). "Visual transformation of size". In: *Journal of Experimental Psychology: Human Perception and Performance* 1.3, pp. 214–220. ISSN: 1939-1277(Electronic);0096-1523(Print). DOI: 10.1037/0096-1523.1.3.214.

Carpenter, Patricia A, Marcel A Just, and Peter Shell (1990). "What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test." In: *Psychological review* 97.3, p. 404.

Chandrasekaran, Balakrishnan et al. (2011). "Augmenting cognitive architectures to support diagrammatic imagination". In: *Topics in cognitive science* 3.4, pp. 760–777.

Chang, Angel X, Manolis Savva, and Christopher D Manning (2014). "Learning Spatial Knowledge for Text to 3D Scene Generation." In: *EMNLP*, pp. 2028–2038.

Cheng, Yi-Ling and Kelly S Mix (2014). "Spatial training improves children's mathematics ability". In: *Journal of Cognition and Development* 15.1, pp. 2–11.

Cohn, Anthony G et al. (1997). "Qualitative spatial representation and reasoning with the region connection calculus". In: *GeoInformatica* 1.3, pp. 275–316.

Cooper, L. A. and R. N. Shepard (1973). "Chronometric studies of the rotation of mental images". In: *Visual Information Processing*. Ed. by W. G. Chase. New York: Academic Press, pp. 75–176.

Coyne, Bob and Richard Sproat (2001). "WordsEye: an automatic text-to-scene conversion system". In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, pp. 487–496.

Croft, David and Paul Thagard (2002). "Dynamic Imagery: A Computational Model of Motion and Visual Analogy". en. In: *Model-Based Reasoning*. Ed. by Lorenzo Magnani and Nancy J. Nersessian. Springer US, pp. 259–274. ISBN: 978-1-4613-5154-2, 978-1-4615-0605-8. URL: http://link.springer.com/chapter/10.1007/978-1-4615-0605-8_15 (visited on 03/01/2015).

Davies, Jim, Ashok K Goel, and Patrick W Yaner (2008). "Proteus: Visuospatial analogy in problem-solving". In: *Knowledge-Based Systems* 21.7, pp. 636–654.

Davis, Ernest (2014). *Representations of commonsense knowledge*. Morgan Kaufmann.

Davis, Randall, Howard Shrobe, and Peter Szolovits (1993). "What is a knowledge representation?" In: *AI magazine* 14.1, p. 17.

De Kleer, Johan and John Seely Brown (1984). "A qualitative physics based on confluences". In: *Artificial intelligence* 24.1-3, pp. 7–83.

DeShon, Richard P., David Chan, and Daniel A. Weissbein (1995). "Verbal overshadowing effects on Raven's advanced progressive matrices: Evidence for multidimensional performance determinants". In: *Intelligence* 21.2, pp. 135–155. ISSN: 0160-2896. DOI: 10.1016/0160-2896(95)90023-3. URL: http://www.sciencedirect.com/science/article/pii/0160289695900233 (visited on 09/24/2014).

Durupınar, Funda, Umut Kahramankaptan, and Ilyas Cicekli (2004). "Intelligent indexing, querying and reconstruction of crime scene photographs". In: *In TAINN*. Citeseer.

Evans, Thomas G (1968). "A program for the solution of geometric-analogy intelligence test questions". In: *Semantic Information Processing*. Ed. by Marvin Minsky. Cambridge, MA: MIT Press, pp. 271–353.

Farah, Martha J and Katherine M Hammond (1988). "Mental rotation and orientation-invariant object recognition: Dissociable processes". In: *Cognition* 29.1, pp. 29–46.

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

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

Ferrucci, David et al. (2010). "Building Watson: An overview of the DeepQA project". In: *AI magazine* 31.3, pp. 59–79.

Finke, Ronald A. and Steven Pinker (1982). "Spontaneous imagery scanning in mental extrapolation". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8.2, pp. 142–147. ISSN: 1939-1285(Electronic);0278-7393(Print). DOI: 10.1037/0278-7393.8.2.142.

Finke, Ronald A., Steven Pinker, and Martha J. Farah (1989). "Reinterpreting Visual Patterns in Mental Imagery". en. In: *Cognitive Science* 13.1, pp. 51–78. ISSN: 1551-6709. DOI: 10.1207/s15516709cog1301_2. URL: http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1301_2/abstract (visited on 05/12/2014).

Finlayson, Mark A. and Patrick H. Winston (2007). "Reasoning by Imagining: The Neo-Bridge System". MIT CSAIL Research Abstract.

Fitzgerald, Tesca et al. (2015). "Visual case retrieval for interpreting skill demonstrations". In: *International Conference on Case-Based Reasoning*. Springer, pp. 119–133.

Földiák, Peter (1991). "Learning Invariance from Transformation Sequences". In: *Neural Computation* 3.2, pp. 194–200.

Forbus, Kenneth D (1984). "Qualitative process theory". In: *Artificial intelligence* 24.1, pp. 85–168.

Forbus, Kenneth et al. (2011). "CogSketch: Sketch understanding for cognitive science research and for education". In: *Topics in Cognitive Science* 3.4, pp. 648–666.

Funt, Brian V. (1980). "Problem-solving with diagrammatic representations". In: *Artificial Intelligence* 13.3, pp. 201–230. ISSN: 0004-3702. DOI: 10.1016/0004-3702(80)90002-8. URL: http://www.sciencedirect.com/science/article/pii/0004370280900028 (visited on 01/29/2015).

Gardin, Francesco and Bernard Meltzer (1989). "Analogical representations of naive physics". In: *Artificial Intelligence* 38.2, pp. 139–159.

Gavrila, Dariu M (1998). "Multi-feature hierarchical template matching using distance transforms". In: *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*. Vol. 1. IEEE, pp. 439–444.

Gelernter, Herbert (1959). "Realization of a geometry theorem proving machine." In: *IFIP Congress*, pp. 273–281.

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

Giunchiglia, Fausto et al. (1992). "Understanding scene descriptions by integrating different sources of knowledge". In: *International journal of man-machine studies* 37.1, pp. 47–81.

Glasgow, Janice, N. Hari Narayanan, and B. Chandrasekaran, eds. (1995). *Diagrammatic Reasoning: Cognitive and Computational Perspectives*. Cambridge, MA, USA: MIT Press. ISBN: 0262571129.

Glasgow, Janice and Dimitri Papadias (1992). "Computational imagery". In: *Cognitive science* 16.3, pp. 355–394.

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

Goebel, R Patrick (1990). "The mathematics of mental rotations". In: *Journal of Mathematical Psychology* 34.4, pp. 435–444.

Goel, Ashok K et al. (1994). "Multistrategy adaptive path planning". In: *IEEE Expert* 9.6, pp. 57–65.

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

Gurr, Corin A (1998). "On the isomorphism, or lack of it, of representations". In: *Visual language theory*, pp. 293–305.

Hamrick, Jessica, Peter Battaglia, and Joshua B Tenenbaum (2011). "Internal physics models guide probabilistic judgments about object dynamics". In: *Proceedings of the 33rd annual conference of the cognitive science society*. Cognitive Science Society Austin, TX, pp. 1545–1550.

Hill, A et al. (1994). "Medical image interpretation: A generic approach using deformable templates". In: *Medical Informatics* 19.1, pp. 47–59.

Hunt, Earl (1974). "Quote the Raven? Nevermore". In: *Knowledge and cognition*. Oxford, England: Lawrence Erlbaum, pp. ix, 321.

Ioerger, Thomas R (1994). "The manipulation of images to handle indeterminacy in spatial reasoning". In: *Cognitive Science* 18.4, pp. 551–593.

Jain, Anil K, Yu Zhong, and Marie-Pierre Dubuisson-Jolly (1998). "Deformable template models: A review". In: *Signal processing* 71.2, pp. 109–129.

Johansson, Richard et al. (2005). "Automatic text-to-scene conversion in the traffic accident domain". In: *IJCAI*. Vol. 5, pp. 1073–1078.

Johnston, BG and Mary-Anne Williams (2009). "Autonomous learning of commonsense simulations". In: *Symposium on Logical Formalizations of Commonsense Reasoning*. UTSe-Press.

Just, Marcel A and Patricia A Carpenter (1985). "Cognitive coordinate systems: accounts of mental rotation and individual differences in spatial ability." In: *Psychological review* 92.2, p. 137.

Kazhdan, Michael, Thomas Funkhouser, and Szymon Rusinkiewicz (2003). "Rotation invariant spherical harmonic representation of 3D shape descriptors". In: *Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. Eurographics Association, pp. 156–164.

Kirby, John R and Michael J Lawson (1983). "Effects of strategy training on progressive matrices performance". In: *Contemporary Educational Psychology* 8.2, pp. 127–140.

Kosslyn, Stephen M. (1973). "Scanning visual images:Some structural implications". en. In: *Perception & Psychophysics* 14.1, pp. 90–94. ISSN: 0031-5117, 1532-5962. DOI: 10.3758/BF03198621. URL: http://link.springer.com/article/10.3758/BF03198621 (visited on 08/13/2014).

Kosslyn, Stephen M., Thomas M. Ball, and Brian J. Reiser (1978). "Visual images preserve metric spatial information: Evidence from studies of image scanning". In: *Journal of Experimental Psychology: Human Perception and Performance* 4.1, pp. 47–60. ISSN: 1939-1277(Electronic);0096-1523(Print). DOI: 10.1037/0096-1523.4.1.47.

Kosslyn, Stephen M, Alvaro Pascual-Leone, et al. (1999). "The role of area 17 in visual imagery: Convergent evidence from PET and rTMS". In: *Science* 284.5411, pp. 167–170.

Kosslyn, Stephen M and Steven P Shwartz (1977). "A simulation of visual imagery". In: *Cognitive Science* 1.3, pp. 265–295.

Kosslyn, Stephen M, William L Thompson, et al. (1995). "Topographical representations of mental images in primary visual cortex". In: *Nature* 378.6556, pp. 496–498.

Kuipers, Benjamin (2000). "The spatial semantic hierarchy". In: *Artificial intelligence* 119.1-2, pp. 191–233.

Kunda, Maithilee (2013). "Visual problem solving in autism, psychometrics, and AI: the case of the Raven's Progressive Matrices intelligence test". PhD thesis. Georgia Tech. URL: https://smartech.gatech.edu/handle/1853/47639 (visited on 09/24/2014).

Kunda, Maithilee and Ashok K. Goel (2011). "Thinking in pictures as a cognitive account of autism". In: *Journal of autism and developmental disorders* 41.9, pp. 1157–1177. URL: http://link.springer.com/article/10.1007/s10803-010-1137-1 (visited on 02/17/2014).

Kunda, Maithilee, Keith McGreggor, and Ashok K. Goel (2013). "A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations". In: *Cognitive Systems Research* 22, pp. 47–66.

Kunda, Maithilee and Julia Ting (2016). "Looking around the mind's eye: Attention-based access to visual search templates in working memory". In: *Advances in cognitive systems* 4, pp. 113–129.

Lakoff, George and Mark Johnson (2008). *Metaphors we live by*. University of Chicago press.

Larkin, Jill H and Herbert A Simon (1987). "Why a diagram is (sometimes) worth ten thousand words". In: *Cognitive science* 11.1, pp. 65–100.

Larsen, Axel and Claus Bundesen (1998). "Effects of spatial separation in visual pattern matching: Evidence on the role of mental translation". In: *Journal of Experimental Psychology: Human Perception and Performance* 24.3, pp. 719–731. ISSN: 1939-1277(Electronic);0096-1523(Print). DOI: 10.1037/0096-1523.24.3.719.

Larsen, Axel, William McIlhagga, and Claus Bundesen (1999). "Visual pattern matching: Effects of size ratio, complexity, and similarity in simultaneous and successive matching". en. In: *Psychological Research* 62.4, pp. 280–288. ISSN: 0340-0727, 1430-2772. DOI: 10.1007/s004260050058. URL: http://link.springer.com/article/10.1007/s004260050058 (visited on 08/19/2014).

Lathrop, Scott D, Samuel Wintermute, and John E Laird (2011). "Exploring the functional advantages of spatial and visual cognition from an architectural perspective". In: *Topics in cognitive science* 3.4, pp. 796–818.

Levesque, Hector J (2014). "On our best behaviour". In: *Artificial Intelligence* 212, pp. 27–35.

Levesque, Hector J, Ernest Davis, and Leora Morgenstern (2011). "The Winograd schema challenge." In: *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*. Vol. 46, p. 47.

Lin, Xiao and Devi Parikh (2015). "Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2984–2993.

Lovett, Andrew et al. (2009). "Solving Geometric Analogy Problems Through Two-Stage Analogical Mapping". In: *Cognitive science* 33.7, pp. 1192–1231.

Luursema, Jan-Maarten, Willem B Verwey, and Remke Burie (2012). "Visuospatial ability factors and performance variables in laparoscopic simulator training". In: *Learning and individual differences* 22.5, pp. 632–638.

Marr, David (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman and Company.

McGreggor, Keith and Ashok Goel (2011). "Finding the odd one out: a fractal analogical approach". In: *Proceedings of the 8th ACM conference on Creativity and cognition*. ACM, pp. 289–298.

McGreggor, Keith and Ashok K Goel (2014). "Confident Reasoning on Raven's Progressive Matrices Tests." In: *AAAI*, pp. 380–386.

McGreggor, Keith, Maithilee Kunda, and Ashok K. Goel (2014). "Fractals and ravens". In: *Artificial Intelligence* 215, pp. 1–23.

Mel, Bartlett W. (1986). "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.

— (1990). *Connectionist Robot Motion Planning*. Academic Press.

Memisevic, Roland and Geoffrey Hinton (2007). "Unsupervised learning of image transformations". In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, pp. 1–8.

Memisevic, Roland and Geoffrey E Hinton (2010). "Learning to represent spatial transformations with factored higher-order boltzmann machines". In: *Neural computation* 22.6, pp. 1473–1492.

Miller, Arthur I (2012). *Insights of genius: Imagery and creativity in science and art*. Springer Science & Business Media.

Moravec, Hans and Alberto Elfes (1985). "High resolution maps from wide angle sonar". In: *Robotics and Automation. Proceedings. 1985 IEEE International Conference on*. Vol. 2. IEEE, pp. 116–121.

Myers, Karen L and Kurt Konolige (1994). "Reasoning with analogical representations". In: *Foundations of Knowledge Representation and Reasoning*. Springer, pp. 229–249.

Narayanan, N Hari and B Chandrasekaran (1991). "Reasoning Visually about Spatial Interactions." In: *IJCAI*, pp. 360–365.

National Research Council (2009). "Mathematics Learning in Early Childhood: Paths toward Excellence and Equity." In: *National Academies Press*.

Nersessian, Nancy J (2008). *Creating scientific concepts*. MIT press.

Newell, Allen and Herbert A Simon (1976). "Computer science as empirical inquiry: Symbols and search". In: *Communications of the ACM* 19.3, pp. 113–126.

Paivio, Allan (2014). *Mind and its evolution: A dual coding theoretical approach*. Psychology Press.

Palmer, Joshua H. and Maithilee Kunda (2018). "Thinking in PolAR Pictures: Using Rotation-Friendly Mental Images to Solve Leiter-R Form Completion". In: *To appear in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Pearson, Joel and Stephen M Kosslyn (2015). "The heterogeneity of mental representation: ending the imagery debate". In: *Proceedings of the National Academy of Sciences* 112.33, pp. 10089–10092.

Perlis, Don (2016). "Five Dimensions of Reasoning in the Wild." In: *AAAI*, pp. 4152–4156.

Petre, Marian and Alan F Blackwell (1999). "Mental imagery in program design and visual programming". In: *International Journal of Human-Computer Studies* 51.1, pp. 7–30.

Polland, Mark J (1996). "Mental Imagery in Creative Problem Solving". PhD thesis. Claremont Graduate School.

Ponzetto, Simone Paolo and Michael Strube (2007). "Deriving a large scale taxonomy from Wikipedia". In: *AAAI*. Vol. 7, pp. 1440–1445.

Prabhakaran, Vivek et al. (1997). "Neural substrates of fluid reasoning: an fMRI study of neocortical activation during performance of the Raven's Progressive Matrices Test". In: *Cognitive psychology* 33.1, pp. 43–63.

Ragni, Marco and Markus Knauff (2013). "A theory and a computational model of spatial reasoning with preferred mental models." In: *Psychological review* 120.3, p. 561.

Rao, Rajesh PN et al. (2002). "Eye movements in iconic visual search". In: *Vision research* 42.11, pp. 1447–1463.

Rasmussen, Daniel and Chris Eliasmith (2011). "A neural model of rule generation in inductive reasoning". In: *Topics in Cognitive Science* 3.1, pp. 140–153.

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

Reisberg, Daniel (2014). *Auditory imagery*. Psychology Press.

Roy, D., Kai-yuh Hsiao, and N. Mavridis (2004). "Mental imagery for a conversational robot". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 34.3, pp. 1374–1383. ISSN: 1083-4419. DOI: 10.1109/TSMCB.2004.823327.

Schirra, Jorg RJ and Eva Stopp (1993). "ANTLIMA: a listener model with mental images". In: *Proceedings of the 13th international joint conference on Artifical intelligence-Volume 1*. Morgan Kaufmann Publishers Inc., pp. 175–180.

Schultheis, Holger and Thomas Barkowsky (2011). "Casimir: an architecture for mental spatial knowledge processing". In: *Topics in cognitive science* 3.4, pp. 778–795.

Schultheis, Holger, Sven Bertel, and Thomas Barkowsky (2014). "Modeling mental spatial reasoning about cardinal directions". In: *Cognitive science* 38.8, pp. 1521–1561.

Schwartz, Daniel L and John B Black (1996). "Analog imagery in mental model reasoning: Depictive models". In: *Cognitive Psychology* 30.2, pp. 154–219.

Seepanomwan, Kristsana et al. (2013). "Modelling mental rotation in cognitive robots". In: *Adaptive Behavior* 21.4, pp. 299–312.

Seversky, Lee M and Lijun Yin (2006). "Real-time automatic 3D scene generation from natural language voice and text descriptions". In: *Proceedings of the 14th ACM international conference on Multimedia*. ACM, pp. 61–64.

Shepard, Roger N. (1984). "Ecological constraints on internal representation: resonant kinematics of perceiving, imagining, thinking, and dreaming". eng. In: *Psychological review* 91.4, pp. 417–447. ISSN: 0033-295X.

Shepard, Roger N and Jacqueline Metzler (1971). "Mental Rotation of Three-Dimensional Objects". In: *Science* 171.3972, pp. 701–703.

Shimojima, Atsushi (1999). "The graphic-linguistic distinction exploring alternatives". In: *Artificial Intelligence Review* 13.4, pp. 313–335.

Shrager, Jeff (1990). "Commonsense perception and the psychology of theory formation". In: *Computational models of scientific discovery and theory formation*, pp. 437–470.

Singh, Push et al. (2002). "Open Mind Common Sense: Knowledge acquisition from the general public". In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, pp. 1223–1237.

Slotnick, Scott D, William L Thompson, and Stephen M Kosslyn (2005). "Visual mental imagery induces retinotopically organized activation of early visual areas". In: *Cerebral cortex* 15.10, pp. 1570–1583.

Snow, Richard E., Patrick C. Kyllonen, and Brian Marshalek (1984). "The topography of ability and learning correlations". In: *Advances in the psychology of human intelligence* 2, pp. 47–103.

Soulières, Isabelle, Michelle Dawson, et al. (2009). "Enhanced visual processing contributes to matrix reasoning in autism". In: *Human brain mapping* 30.12, pp. 4082–4107.

Soulières, Isabelle, T A Zeffiro, et al. (2011). "Enhanced mental image mapping in autism". eng. In: *Neuropsychologia* 49.5, pp. 848–857. ISSN: 1873-3514. DOI: 10.1016/j.neuropsychologia.2011.01.027.

Steels, Luc (1988). "Steps towards common sense". In: *Proceedings of the 8th European Conference on Artificial Intelligence (ECAI)*.

Stevenson, Richard J and Trevor I Case (2005). "Olfactory imagery: a review". In: *Psychonomic Bulletin & Review* 12.2, pp. 244–264.

Strannegård, Claes, Simone Cirillo, and Victor Ström (2013). "An anthropomorphic method for progressive matrix problems". In: *Cognitive Systems Research* 22, pp. 35–46.

Tabachneck-Schijf, Hermina JM, Anthony M Leonardo, and Herbert A Simon (1997). "CaMeRa: A computational model of multiple representations". In: *Cognitive Science* 21.3, pp. 305–350.

Tarr, Michael J and Steven Pinker (1989). "Mental rotation and orientation-dependence in shape recognition". In: *Cognitive psychology* 21.2, pp. 233–282.

Thagard, Paul (1996). *Mind: Introduction to cognitive science*. Vol. 4. MIT press Cambridge, MA.

Uttal, David H et al. (2013). "The malleability of spatial skills: a meta-analysis of training studies." In: *Psychological bulletin* 139.2, p. 352.

Vanrie, Jan, Erik Béatse, et al. (2002). "Mental rotation versus invariant features in object perception from different viewpoints: An fMRI study". In: *Neuropsychologia* 40.7, pp. 917–930.

Vanrie, Jan, Bert Willems, and Johan Wagemans (2001). "Multiple routes to object matching from different viewpoints: Mental rotation versus invariant features". In: *Perception* 30.9, pp. 1047–1056.

Wai, Jonathan, David Lubinski, and Camilla P Benbow (2009). "Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance." In: *Journal of Educational Psychology* 101.4, p. 817.

Waltz, David L and Lois Boggess (1979). *Visual Analog Representations for Natural Language Understanding.* Tech. rep. DTIC Document.

Winston, Patrick H. (1992). *Artificial Intelligence*.

Winston, Patrick Henry (2012). "The right way". In: *Advances in Cognitive Systems* 1, pp. 23–36.

Yaner, Patrick W and Ashok K Goel (2008). "Analogical recognition of shape and structure in design drawings". In: *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 22.02, pp. 117–128.

Yoo, Seung-Schik et al. (2003). "Neural substrates of tactile imagery: a functional MRI study". In: *Neuroreport* 14.4, pp. 581–585.

Yuille, Alan L, Peter W Hallinan, and David S Cohen (1992). "Feature extraction from faces using deformable templates". In: *International journal of computer vision* 8.2, pp. 99–111.

Zacks, Jeffrey M (2008). "Neuroimaging studies of mental rotation: a meta-analysis and review". eng. In: *Journal of cognitive neuroscience* 20.1, pp. 1–19. ISSN: 0898-929X. DOI: 10.1162/jocn.2008.20013.

Zelinsky, Gregory J (2008). "A theory of eye movements during target acquisition." In: *Psychological review* 115.4, p. 787.

Zeman, A, M Dewar, and S Della Sala (2015). "Lives without imagery-Congenital aphantasia." In: *Cortex* 73, pp. 378–380.