COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 35 (2011) 939–962

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2011.01176.x

Are People Successful at Learning Sequences of Actions on a Perceptual Matching Task?

Reiko Yakushijin,[a] Robert A. Jacobs[b]

[a]*Department of Psychology, Aoyama Gakuin University, Tokyo*
[b]*Department of Brain & Cognitive Sciences, University of Rochester*

Received 5 October 2009; received in revised form 8 November 2010; accepted 30 November 2010

Abstract

We report the results of an experiment in which human subjects were trained to perform a perceptual matching task. Subjects were asked to manipulate comparison objects until they matched target objects using the fewest manipulations possible. An unusual feature of the experimental task is that efficient performance requires an understanding of the hidden or latent causal structure governing the relationships between actions and perceptual outcomes. We use two benchmarks to evaluate the quality of subjects' learning. One benchmark is based on optimal performance as calculated by a dynamic programming procedure. The other is based on an adaptive computational agent that uses a reinforcement-learning method known as Q-learning to learn to perform the task. Our analyses suggest that subjects were successful learners. In particular, they learned to perform the perceptual matching task in a near-optimal manner (i.e., using a small number of manipulations) at the end of training. Subjects were able to achieve near-optimal performance because they learned, at least partially, the causal structure underlying the task. In addition, subjects' performances were broadly consistent with those of model-based reinforcement-learning agents that built and used internal models of how their actions influenced the external environment. We hypothesize that people will achieve near-optimal performances on tasks requiring sequences of action—especially sensorimotor tasks with underlying latent causal structures—when they can detect the effects of their actions on the environment, and when they can represent and reason about these effects using an internal mental model.

Keywords: Sequential action; Action learning; Perceptual matching; Ideal actor; Reinforcement learning; Causal learning

Correspondence should be sent to Reiko Yakushijin, Department of Psychology, Aoyama Gakuin University, Shibuya-ka, 150-8366, Japan. E-mail: yaku@ephs.aoyama.ac.jp

1. Introduction

Tasks requiring people to make a sequence of actions to reach a goal are commonplace in our lives. When playing chess, a person must make a sequence of chess moves to capture an opponent's king. When driving to work, a person must make a sequence of left and right turns to arrive at work in a timely manner. When pursuing health goals, a person must make a sequence of food and exercise choices to reach a desired weight. And when pursuing financial goals, a person must make a sequence of saving and spending choices to achieve a financial target. Unsurprisingly, interest in sequential action tasks among cognitive scientists has increased dramatically in recent years (e.g., Busemeyer, 2001; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Fu & Anderson, 2006; Gibson, Fichman, & Plaut, 1997; Gonzalez, Vanyukov, & Martin, 2005; Gureckis & Love, 2009a,b; Shanks, Tunney, & McCarthy, 2002; Stanley, Mathews, Buss, & Kotler-Cope, 1989; Sutton & Barto, 1998).

Here, we focus on a particular type of sequential action task known as a perceptual matching task. Perceptual matching tasks are commonplace in our everyday lives. They occur when we attempt to mimic some aspect of our environments. For example, when a student in a dance class learns from an instructor, the student makes a sequence of actions (chosen from the set of possible arm, leg, and body movements) so that the student's movements mimic those of the instructor. Or when an artist draws a person, object, or scene, the artist makes a sequence of actions (chosen from the set of possible pen colors and pen strokes) so that the drawing resembles the visual target. Or when a person cooks a dish previously eaten in a restaurant, the person makes a sequence of actions (chosen from the set of possible ingredients and ways of combining them) so that the food resembles the food served in the restaurant. Perceptual matching tasks have previously been used to study sequential actions in the contexts of visual cognition and sensorimotor control (e.g., Ballard, Hayhoe, Pook, & Rao, 1997; Gray, Sims, Fu, & Schoelles, 2006).

An important aspect of the perceptual matching task used in the experiment reported here is that efficient performance requires knowledge (possibly implicit) of a hidden or latent causal structure governing the relationships between actions and perceptual outcomes. In many everyday tasks, people are required to understand the interactions, or ''causal relations,'' among actions and their effects (Busemeyer, 2001; Gopnik & Shulz, 2007). For example, when a student in a dance class moves his or her arm to mimic the instructor's arm movement, the student must understand that forces exerted at the shoulder also influence the positions and velocities of the elbow, wrist, and fingers. To make an efficient movement, the student must use this knowledge of the causal interactions among forces and motor states to design an effective motor plan. The student does not directly receive instruction regarding these causal interactions, and thus, the student must acquire knowledge of these interactions in a trial-and-error manner while learning to dance. Our experiment mimicked this type of situation in the sense that subjects were required to learn about latent causal interactions in a trial-and-error manner while performing the experimental task.

Our primary interest is in whether people are successful at learning to perform a sequential action task—specifically a perceptual matching task—with an underlying latent causal

structure. Methodologically, we evaluate the quality of our subjects' learning in two different ways. These ways differ in terms of the benchmarks to which subjects' performances are compared. One way uses a benchmark of optimal performance on a task. Analyses based on optimal performance are referred to as ideal observer analyses, ideal actor analyses, or rational analyses in the literatures on perception, motor control, and cognition, respectively (e.g., Anderson, 1990; Geisler, 2004; Todorov, 2004). At each moment during training with a task, a learner's performance can be compared to the optimal performance for that task. If a learner achieves near-optimal performance at the end of training, then it can be claimed that the learner has been successful.

This general approach has previously been used to study the performances—though not necessarily the learning—of different organisms on different sequential action tasks. In some instances, researchers have found that organisms perform in a near-optimal manner either at the end of laboratory training or in the absence of training. Stephens and Krebs (1986) argued that many species forage in a near-optimal manner, and that this optimality helps explain many of their behaviors. Lee (2006) found that at least some people performed near-optimally on a ''stopping'' problem in which people are sequentially presented with alternatives and it is efficient to make a selection without waiting to see all alternatives. In addition, he found no evidence for learning during the course of the experiment. Chhabra and Jacobs (2006) found that people performed near-optimally at the end of training on an adaptive control task in a variety of noise environments. In contrast to Lee (2006), however, they found evidence for learning during the experiment.

In other instances, it seems that people perform in a suboptimal manner, even at the end of extensive laboratory training. Stankiewicz, Legge, Mansfield, and Schlicht (2006) reported that people performed suboptimally on a navigation task due to the fact that they often confused spatial locations with identical visual appearances (a phenomenon known as ''perceptual aliasing''; McCallum, 1993; Whitehead & Ballard, 1991). Neth, Sims, and Gray (2005, 2006) and Gureckis and Love (2009a,b) found that people often performed suboptimally on a task in which it is efficient to choose options that produce significant long-term gains despite minimal short-term gains.

In addition to comparing a learner's performances on our experimental task with the optimal performance on the task, we also analyze our data by comparing a subject's learning performances with those of adaptive computational agents that are trained to perform the same task. We consider agents that learn via ''reinforcement-learning'' methods developed by researchers interested in artificial intelligence (Sutton & Barto, 1998). Cognitive scientists have begun to use reinforcement-learning methods to develop new theories of biological learning (Busemeyer & Pleskac, 2009; Daw & Touretzky, 2002; Fu & Anderson, 2006; Schultz, Dayan, & Montague, 1997; Sun, Slusarz, & Terry, 2005). Because reinforcement learning is regarded as effective and well-understood from an engineering perspective, and as plausible from psychological and neurophysiological perspectives, the performances of agents based on this form of learning can provide useful benchmarks for evaluating a person's learning. If a person's performances during training improve at the same rate as those of a reinforcement-learning agent, then it can be argued that the person is a successful learner. If a person's performances improve at a slower rate than those of the agent, then the

person is not learning as much from experience as he or she could learn. Experimentation is often required to identify the cognitive ''bottlenecks'' preventing the person from learning faster. Lastly, if a person's performances improve at a faster rate, then this might suggest, for example, that the person is using information sources that are not available to the agent. A new, more complex agent should be considered in this case.

The research reported in this article is unusual for several reasons. Although earlier articles have considered sequential action tasks, there are relatively few articles that have considered perceptual matching tasks and even fewer (none to our knowledge) that have considered perceptual matching tasks in which relationships between actions and outcomes are governed by latent causal structures. In addition, as mentioned above, we combine our experimental study with sophisticated computational analyses. We compare human performances on the experimental task to optimal performance on the task, where optimal performance is calculated via dynamic programming. To our knowledge, only one other paper has compared human performances with optimal performance on a sequential action task as calculated by dynamic programming (Chhabra & Jacobs, 2006). We also compare human learning on the experimental task with the learning performances of model-free and model-based reinforcement-learning agents. To date, detailed comparisons of learning performances between humans and reinforcement-learning agents are relatively rare in the scientific literature (e.g., Fu & Anderson, 2006; Gray et al., 2006).

In summary, we analyze our experimental data by comparing subjects' learning performances to optimal performance on the experimental task and to the performances of adaptive computational agents. These comparisons suggest that subjects were successful learners. In particular, subjects learned to perform the perceptual matching task in a near-optimal manner at the end of training. Subjects were able to achieve near-optimal performance because they learned, at least partially, the causal structure underlying the task. In addition, subjects' performances were broadly consistent with those of model-based reinforcement-learning agents that built and used internal models of how their actions influenced the external environment. We hypothesize that people will achieve near-optimal performances on sequential action tasks—especially sensorimotor tasks with latent causal structures—when they can detect the effects of their actions on the environment, and when they can represent and reason about these effects using an internal mental model.

2. Experiment

The perceptual matching task used visual objects from a class of parameterized objects known as ''supershapes'' (Gielis, 2003). The parameters were hidden or latent variables whose values determined the shapes of the objects. On each trial, subjects viewed a target object, a comparison object, and a set of six buttons. The buttons were organized into three pairs, and each pair could be used to decrease or increase the value of an action variable. By pressing the buttons, subjects could change the values of the action variables which, in turn, changed the values of the parameters underlying the comparison object's shape which, in

turn, changed the shape of the comparison object. Subjects' task was to press one or more buttons (i.e., to change the values of the action variables) to modify the shape of the comparison object until it matched the shape of the target object using as few button presses as possible.

A specific experimental condition was characterized by a specific set of causal relations among the latent shape parameters. For example, one such set is schematically illustrated in Fig. 1. Here, the three action variables are denoted *A*, *B*, and *C*. These variables are observable in the sense that subjects could directly and easily control their values through the use of the buttons. The values of the action variables determined the values of the shape parameters, denoted *X*, *Y*, and *Z*. Note that there are causal relations among the shape parameters. According to the network in Fig. 1, if the value of *X* is changed, then this leads to a modification of *Y* which, in turn, leads to a modification of *Z*. The shape parameters determine the shape of the comparison object, whose perceptual features are denoted f_1, f_2, f_3, f_4, f_5, and f_6. The perceptual features used by a subject to assess the similarity of target and comparison object shapes may only be implicitly known by a subject, and their number and nature may differ between subjects.

Importantly, to efficiently convert the comparison object's shape to the target object's shape (i.e., with the fewest number of button presses) often requires an understanding of the causal relations among the shape parameters. For instance, if the values of parameters *X*, *Y*, and *Z* all need to be modified, a person who does not understand the causal relations among the shape parameters may decide to change the value of action variable *C* (thereby changing shape parameter *Z*), then the value of action variable *B* (thereby changing *Y* and *Z*), and finally the value of action variable *A* (thereby changing *X*, *Y*, and *Z*). In many cases, this will be an inefficient strategy. A person with good knowledge of the causal relations among the shape parameters knows that he or she can change the values of *X*, *Y*, and *Z* with a single button press that decreases or increases the value of action variable *A*. Thus, a good understanding of the causal relations among the shape parameters will lead to efficient task performance; however, a poor understanding of the causal relations will lead to many more button presses than necessary.

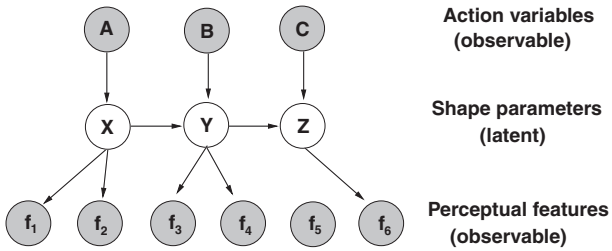

Fig. 1. A Bayesian network representing the causal relations (in one of the experimental conditions) among the observable action variables *A*, *B*, and *C*, the latent shape parameters *X*, *Y*, and *Z*, and the observable perceptual features f_1, f_2, f_3, f_4, f_5, and f_6. For the sake of simplicity, this network does not represent the fact that subjects' button presses determined the values of the action variables.

2.1. Methods

2.1.1. Subjects

Twenty-four undergraduate students at the University of Rochester participated in the experiment. Subjects were paid $10 for their participation. All subjects had normal or corrected-to-normal vision.

2.1.2. Stimuli

Visual stimuli depicted three-dimensional objects whose shapes belonged to a parameterized family of shapes known as ''supershapes'' (Gielis, 2003). In polar coordinates, the radii of the objects in horizontal and vertical dimensions, denoted r_1 and r_2 respectively, are given by the equations:

$$r_1(\theta) = \left[\left| \frac{\cos\left(\frac{m_1\theta}{4}\right)}{a_1} \right|^{n_{12}} + \left| \frac{\sin\left(\frac{m_1\theta}{4}\right)}{b_1} \right|^{n_{13}} \right]^{-\frac{1}{n_{11}}}$$

$$r_2(\phi) = \left[\left| \frac{\cos\left(\frac{m_2\phi}{4}\right)}{a_2} \right|^{n_{22}} + \left| \frac{\sin\left(\frac{m_2\phi}{4}\right)}{b_2} \right|^{n_{23}} \right]^{-\frac{1}{n_{21}}}$$

where θ $(-\pi < \theta < \pi)$ and φ $\left(-\frac{\pi}{2} < \phi < \frac{\pi}{2}\right)$ index angles in the horizontal and vertical dimensions. Polar coordinates are converted to Cartesian coordinates using the equations:

$$x = r_1(\theta)\cos(\theta) r_2(\phi)\cos(\phi)$$

$$y = r_1(\theta)\sin(\theta) r_2(\phi)\cos(\phi)$$

$$z = r_2(\phi)\sin(\phi)$$

Supershapes contain 12 free parameters: m_1, m_2, a_1, a_2, b_1, b_2, n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, and n_{23}. In the experiment, the values of four supershape parameters were linked to the three shape parameters X, Y, and Z as follows: $m_1 = X$, $n_{12} = n_{13} = Y$, and $m_2 = Z$. The values of all other supershape parameters were set to 1.

Fig. 2 illustrates some of the shapes used in the experiment. The top row shows a shape, along with how the shape is modified as shape parameter X increases in value. The middle and bottom rows show how a shape changes as parameters Y and Z increase in value, respectively.

Training and test trials used different shapes for the target objects. On each training trial, a target object was formed by setting its shape parameters X, Y, and Z to either (5, 7, 5), (5, 8, 6), (6, 7, 6), or (6, 8, 5). On each test trial, these values were set to either (5, 7, 6), (5, 8, 5), (6, 7, 5), or (6, 8, 6). On both training and test trials, a comparison object was formed by initializing its shape parameters to values that were within one integer unit

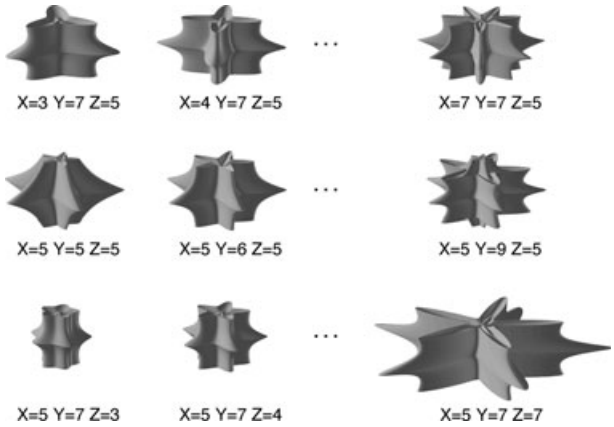

Fig. 2. Examples of shapes used in the experiments. The top, middle, and bottom rows illustrate how the shapes change when the shape parameters *X*, *Y*, and *Z* increase their values.

of those of the target object. Let X_t, Y_t, and Z_t denote the values of a target object's shape parameters, and let X_c, Y_c, and Z_c denote the initial values of a comparison object's shape parameters. Then $X_c \in \{X_t - 1, X_t, X_t + 1\}$, $Y_c \in \{Y_t - 1, Y_t, Y_t + 1\}$, and $Z_c \in \{Z_t - 1, Z_t, Z_t + 1\}$. The sole exception was that a comparison object's shape was never initialized to be equal to the target object's shape. In other words, the shape of the comparison object was initialized to be a perturbation of the shape of the target object, and there were 26 possible perturbations.

2.1.3 Procedure

Subjects performed the experiment in a small, darkened room. Computer displays were presented on a 21-inch CRT monitor whose resolution (in pixels) was set to 1600×1200. Subjects viewed the displays from a distance of approximately 60 cm. At this distance, the image of a target object subtended approximately 5° of visual angle in each of the horizontal and vertical dimensions. The image of a comparison object ranged from 2° to 14° in the horizontal dimension and 3° to 8° in the vertical dimension. Each subject participated in a single experimental session lasting about an hour.

The experiment included six experimental conditions differing in the causal relations among the shape parameters *X*, *Y*, and *Z*. The six possible causal relations are shown in Fig. 3. Two of the causal relations are ''linear'' structures (one parameter has a direct causal influence on a second parameter, which, in turn, has a direct causal influence on a third parameter), two of the relations are ''common cause'' structures (one parameter has direct causal influences on the two remaining parameters), and two of the relations are ''common effect'' structures (two parameters have direct causal influences on a third parameter). Subjects were randomly assigned to one of the six experimental conditions. Each condition included both training and test trials.

On a training trial, a target object shape and initial comparison object shape were randomly selected as described above. The top portion of Fig. 4 shows a typical display at the

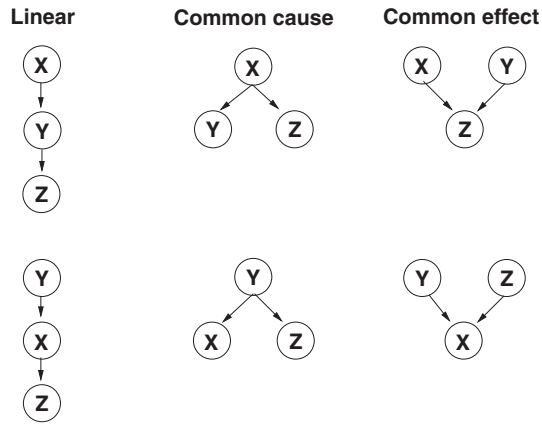

Fig. 3. Bayesian networks illustrating the six causal relations used in Experiment 1. The networks in the left column represent linear structures, the networks in the middle column represent common cause structures, and the networks in the right column represent common effect structures.

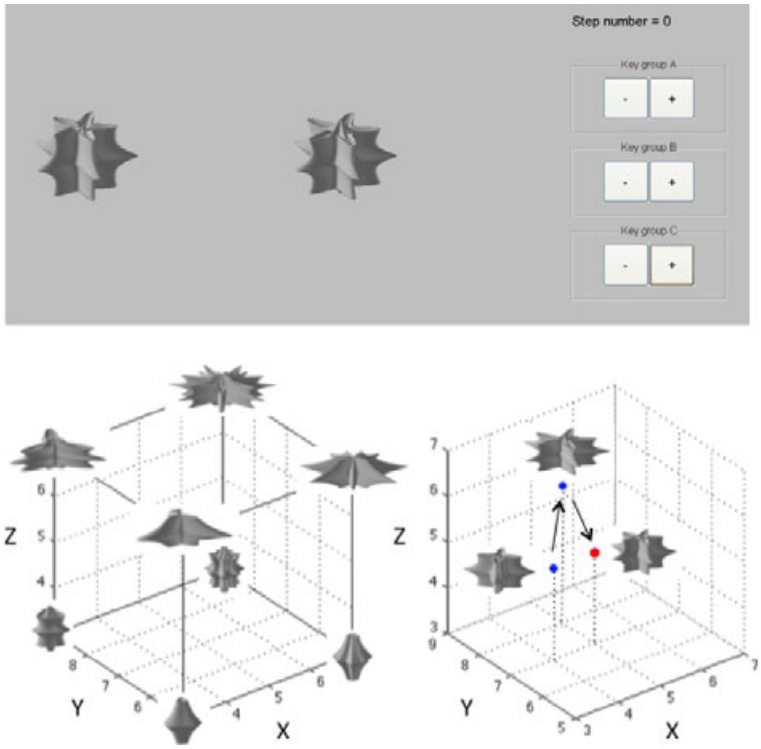

Fig. 4. (Top) An example of a grayscale display used in Experiment 1. A target object is on the left side of the display, a comparison object is in the middle, and the three pairs of buttons used to decrease or increase the values of the action variables *A*, *B*, and *C* are on the right. (Bottom) Illustrations of the ''shape space'' defined by the shape parameters *X*, *Y*, and *Z*. See text for explanation.

start of a training trial. The target object is on the left side of the display, the comparison object is in the middle, and the three pairs of buttons used to decrease or increase the values of action variables A, B, and C are on the right. A subject's task was to change the values of the action variables by pressing the buttons so as to convert the comparison object shape into the target object shape using as few button presses as possible. The step number displayed above the buttons was the number of button presses that the subject had executed during the current trial. When the subject successfully converted the comparison object shape to the target object shape, a high-pitched sound was presented. If a subject pressed the buttons 10 times without converting the comparison object shape to the target object shape, a low-pitched sound was presented and the trial was terminated.[1] If a subject pressed a button which would have moved a shape parameter value out of its allowable range (i.e., more than three units away from the corresponding value of the target object), the button press was ignored and an error message appeared stating that the button press was not currently allowed.

To further illustrate the experimental task, the bottom portion of Fig. 4 illustrates the ''shape space'' defined by the latent shape parameters. The eight shapes in the bottom, left panel are the shapes at the corners of a three-dimensional cube whose sides each have a length of four units, and which is centered at $X = 5$, $Y = 7$, and $Z = 5$. To understand the bottom, right panel, consider a subject trained with the linear causal structure $X \rightarrow Y \rightarrow Z$ (as in Fig. 1). Suppose that at some moment in time, the comparison object has shape parameters $X = 4$, $Y = 7$, and $Z = 5$. Then the subject presses a button to increase the value of action variable A, thereby increasing the values of shape parameters X, Y, and Z (meaning, the parameters now have values $X = 5$, $Y = 8$, and $Z = 6$). Finally, the subject presses a button to decrease the value of action variable B, thereby decreasing the values of shape parameters Y and Z (now, $X = 5$, $Y = 7$, and $Z = 5$). This sequence of three comparison object shapes is illustrated in the bottom, right panel.

Test trials tested subjects' one-step look-ahead knowledge. They were similar to training trials with the following exceptions. On a test trial, subjects had to decide if the comparison object shape could be converted to the target object shape using a single button press. If so, subjects were instructed to press the appropriate button. If not, subjects pressed a button labeled ''Discard.'' Subjects did not receive feedback on test trials. That is, they were not informed about the correctness of their responses, and the comparison object did not change shape if subjects pressed one of the six buttons controlling the action variables.

An experimental session consisted of seven blocks of trials where a block contained a set of training trials followed by a set of test trials. Each set contained 26 trials, one trial for each possible perturbation of a target object shape to form an initial comparison object shape.

2.2. Results

We first report data on subjects' task performances including comparisons between their performances and optimal task performance, then report data regarding subjects' understandings of the causal relations among shape parameters, and finally report data comparing

subjects' learning curves with those of adaptive computational agents that learn via reinforcement-learning methods.

2.2.1. Task performances

As a benchmark for evaluating subjects' performances on the training trials, we first computed the optimal performances on these trials in the six experimental conditions using an optimization method known as dynamic programming (Bellman, 1957; Cormen, Leiserson, Rivest, & Stein, 2001). Recall that for every condition, a set of training trials contained 26 trials corresponding to 26 possible perturbations of the target object to form the initial shape of the comparison object. For the experimental conditions in which the causal relations among the shape parameters was a linear structure or a common effect structure, six of the 26 trials could optimally be performed in one step (i.e., one button press), six could be performed in two steps, ten could be performed in three steps, two could be performed in four steps, and four could be performed in five steps. For the experimental conditions in which the causal relations was a common cause structure, six trials could optimally be performed in one step, eight could be performed in two steps, six could be performed in three steps, four could be performed in four steps, and two could be performed in five steps. For all conditions, the average optimal number of steps was 2.54. Thus, the experimental conditions were well balanced in terms of their intrinsic difficulties.

The three graphs in Fig. 5 show subjects' average learning curves on the sets of training trials for experimental conditions with linear, common cause, and common effect causal structures, respectively. The horizontal axis of each graph shows the block number, and the vertical axis shows the average difference between the number of steps (i.e., button presses) used by subjects during a trial and the optimal number of steps for that trial as computed by

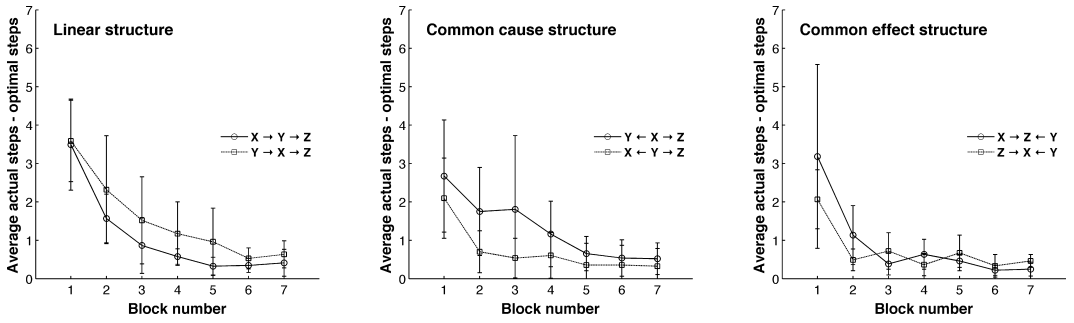

Fig. 5. The three graphs show subjects' learning performances on the training trials for experimental conditions with linear, common cause, and common effect causal structures, respectively. The horizontal axis of each graph shows the block number, and the vertical axis shows the average difference between the number of steps (i.e., button presses) used by subjects during a trial and the optimal number of steps for that trial as computed by the dynamic programming procedure. (For each subject, the average difference for a block of trials was computed. These values were then averaged across subjects. Error bars indicate the standard deviations across subjects.) The solid and dotted lines in each graph plot the data for the subjects in the two experimental conditions using that graph's causal structure.

the dynamic programming procedure. (For each subject, the average difference for a block of trials was computed. These values were then averaged across subjects. Error bars indicate the standard deviations across subjects.) The solid and dotted lines in each graph plot the data for the subjects in the two experimental conditions using that graph's causal structure. These graphs show a number of important features of subjects' performances. First, subjects often found the task to be difficult toward the start of the experiment and, thus, their performances were highly suboptimal during this time period. This poor performance is consistent with subjects' verbal reports that they felt as if they were pressing buttons in a near-random manner during the initial blocks of the experiment.[2] Second, subjects learned during the course of the experiment. Third, subjects achieved near-optimal performances at the end of training: The average difference between a subject's performance and the optimal performance at the end of training is less than half of a step ($M = 0.434$; $SD = 0.324$).

We applied a mixed-design ANOVA (three causal structures as the between-subjects variable × seven training blocks as the within-subjects variable) to the data in Fig. 5. There was a significant effect of block number ($F(6, 126) = 47.41$, $p < .001$), indicating that subjects' performances improved as they received more training. The main effect of causal structure ($F(2, 21) = 1.70$, $p = .207$) and the interaction between causal structure and block number ($F(12, 126) = 1.72$, $p = .067$) were not statistically significant, although the F-value for the interaction was close to significant. Hence, we cannot conclude that subjects' performances differed when different causal structures were used.

In regard to test trials, the three graphs in Fig. 6 show subjects' average percent corrects on these trials for the experimental conditions with linear, common cause, and common effect causal structures, respectively. The horizontal axis of each graph gives the block number, and the vertical axis gives the average percent correct. Based on a mixed-design ANOVA (three causal structures as the between-subjects variable × seven training blocks as the within-subjects variable), there was a significant effect of block number ($F(6, 126) = 10.33$, $p < .001$), suggesting that subjects' performances improved on the test trials as they

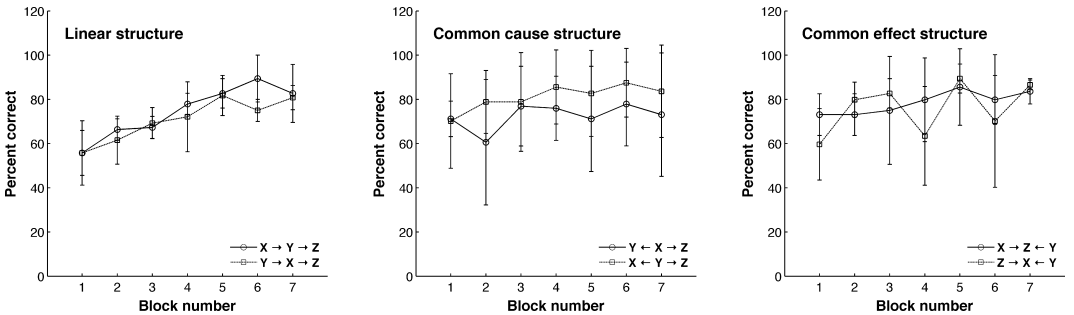

Fig. 6. The three graphs show subjects' average percent corrects on the test trials for experimental conditions with linear, common cause, and common effect causal structures, respectively. The horizontal axis of each graph gives the block number, and the vertical axis gives the average percent correct (error bars indicate the standard deviations across subjects, as in Fig. 5). The solid and dotted lines in each graph plot the data for the subjects in the two experimental conditions using that graph's causal structure.

received more training. Although the main effect of causal structure was not significant ($F(2, 21) = 0.234$, $p = .793$), the interaction between causal structure and block number was significant ($F(12,126) = 2.05$, $p = .025$). Paired comparisons between causal structures at each block show that performances with the linear structure were significantly lower than with the common cause structure only at the first block ($p < .05$). That is, after the first block of training, differences in performances on the test trials in different causal conditions were not statistically significant.

Taken as a whole, data from both training and test trials suggest that subjects improved in their task performances during the course of the experiment. Indeed, data from the training trials show that subjects achieved near-optimal performances. These results are consistent with the idea that subjects learned about the causal relations among shape parameters. To more directly evaluate this idea, we performed additional analyses.

2.2.2. Causal learning

To assess whether data from the training trials support the conclusion that subjects showed good causal learning, we examined the order in which subjects pressed buttons during these trials. The order in which a subject button-pressed should reflect something about the subject's understanding of the causal relations among the shape parameters (even if only imperfectly). For example, consider a subject trained with a linear structure in which X has a direct causal influence on Y which, in turn, has a direct causal influence on Z ($X \rightarrow Y \rightarrow Z$). If the subject wants to change the values of X, Y, and Z, then the subject should press the buttons that modify the value of action variable A (see Fig. 1). If the subject wants to change Y and Z, then the subject should press buttons so as to modify action variable B, and if the subject only wants to change Z, then the subject should press buttons so as to modify C. Consequently, the action variables can be assigned a hierarchical ordering in which A is superordinate to B, which, in turn, is superordinate to C. For the purposes of this analysis, we reasoned that if a subject partially understands the causal relations among the shape parameters and wants to perform the task using the fewest number of button presses, we should expect to see many instances in which a subject modifies a superordinate action variable before modifying a subordinate variable.[3]

We examined the orderings of subjects' button presses by measuring the rate at which consecutive button presses modified a superordinate action variable after a subordinate variable. This rate is referred to as the ''reversed-order rate.'' For example, suppose that a subject trained with the linear structure $X \rightarrow Y \rightarrow Z$ pressed seven buttons on a trial, thereby modifying the action variables in the following order: *ABACCBA*. In this sequence, there are three neighboring pairs of variables in which a superordinate variable was modified after a subordinate variable and, thus, the reversed-order rate is $3/6 = 0.5$. A large reversed-order rate indicates that the subject did not understand the causal relations among the shape parameters, whereas a small rate suggests that the subject did understand these relations.

The three graphs in Fig. 7 show subjects' reversed-order rates for the experimental conditions with linear, common cause, and common effect causal structures, respectively. The horizontal axis of each graph gives the block number, and the vertical axis gives the average reversed-order rate. The two solid lines in each graph plot the data for the subjects in the

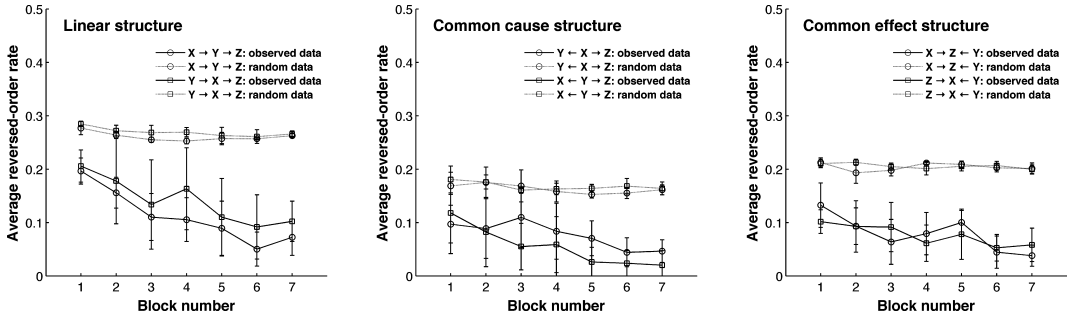

Fig. 7. The three graphs show subjects' average reversed-order rates for experimental conditions with linear, common cause, and common effect causal structures, respectively. The horizontal axis of each graph gives the block number, and the vertical axis gives the average reversed-order rate. The two solid lines in each graph plot the data for the subjects in the two experimental conditions using that graph's causal structure. The two dotted lines show the reversed-order rates of a simulated agent that pressed buttons (i.e., modified action variables) at random.

two experimental conditions using that graph's causal structure. The two dotted lines show the reversed-order rates of a simulated agent that pressed buttons (i.e., modified action variables) at random. This agent formed all action sequences of the same lengths as the subjects and with the same components, but in a random order. (If a subject changed action variables *ABACCBA* on a trial, then the agent formed all sequences of length seven containing three *A*s, two *B*s, and two *C*s.) Based on paired *t* tests, subjects' rates were significantly lower than those of the simulated agent on every block in every causal structure (all *t* tests were significant at the $p < .05$ level). In addition, subjects' rates were always significantly lower in the last two blocks of the experiment than in the first two blocks ($p < .05$). These results suggest that subjects had at least partial knowledge of the causal relations among the shape parameters, and that their causal knowledge increased during the course of training.

An analysis of subjects' understandings of the effects of pressing buttons on test trials provides additional evidence that subjects learned about the causal relations among shape parameters. This analysis was limited to test trials in which a subject pressed one of the six buttons modifying the action variables. We measured the correlations between a subject's decision to modify an action variable (ignoring whether the variable was decreased or increased) and the absolute value of the difference between the target and comparison objects along each shape parameter.

For example, consider button presses modifying action variable *B*. Suppose that every time the target and comparison objects differed in their values of shape parameter *X*, a subject never pressed a button modifying variable *B*. However, if the target and comparison objects differed in their values of *Y*, the subject often pressed a button modifying *B* and, similarly, if the objects differed in their values of *Z*, the subject again often pressed a button modifying *B*. In this case, there would be a negative correlation between modifications of action variable *B* and differences in shape parameter *X*, and positive correlations between modifications of *B* and differences in *Y* and in *Z*. Based on these correlations, we can

conclude that the subject believed that action variable *B* did not influence shape parameter *X*, but that it did influence shape parameters *Y* and *Z*.

The results of this analysis are shown in Fig. 8. The diagrams in the top, middle, and bottom rows correspond to experimental conditions with linear, common cause, and common

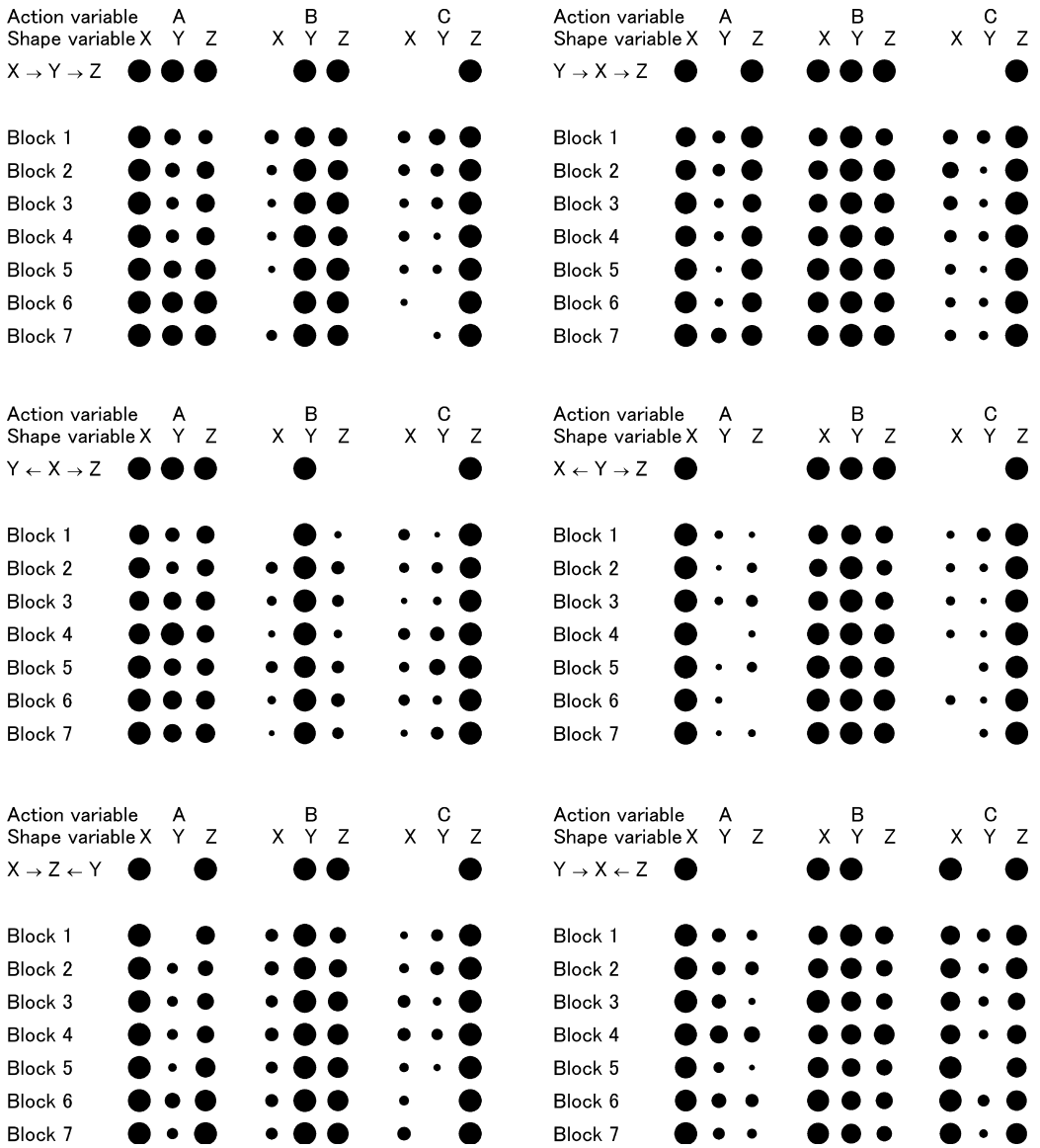

Fig. 8. The average relative correlations between subjects' decisions to modify an action variable and the absolute value of the difference between the target and comparison objects along each shape parameter. See text for explanation.

effect causal structures, respectively. In each diagram, the top portion gives the labels for the action variables and shape parameters. The seven rows in the bottom portion correspond to the seven blocks in an experimental session. The diameter of a circle is proportional to the relative magnitude of an average correlation. Remarkably, these diagrams show that subjects developed excellent understandings of the relationships between action variables and shape parameters. For instance, consider the top left diagram in Fig. 8. Subjects whose data are plotted in this diagram were trained with the linear causal structure $X \rightarrow Y \rightarrow Z$. The correlations plotted in this diagram show that, by the end of training, subjects clearly understood that modifications of action variable A influenced the values of shape parameters X, Y, and Z, modifications of B influenced Y and Z, and modifications of C influenced Z. As a second example, consider the middle right diagram in Fig. 8. For this diagram, subjects were trained with the common cause structure $X \leftarrow Y \rightarrow Z$. By the end of training, subjects understood that modifications of action variable B influenced shape parameters X, Y, and Z, modifications of A influenced X, and modifications of C influenced Z. Based on this data, it appears that subjects developed good understandings of the causal relations characterizing their experimental conditions.

2.2.3. Interim summary

The experiment studied subjects' abilities to learn to perform a perceptual matching task when the task included causal relations among latent parameters governing objects' shapes. We found that, by the end of training, subjects performed in a near-optimal manner. An examination of the sequential order of subjects' actions on the training trials, and an examination of correlations between actions and differences between objects' shapes on test trials, both provided suggestive evidence that subjects learned about the causal relations among the latent shape parameters. We hypothesize that subjects learned the causal relations among the shape parameters (at least partially), and that this causal knowledge underlay their near-optimal task performances.

Above, our analysis of subjects' data used a benchmark of optimal performance based on an optimization technique known as dynamic programming. Although very useful, this analysis does not allow us to evaluate the quality of subjects' rates of learning. To do so, we use a different benchmark based on an adaptive computational agent that uses a reinforcement-learning method known as Q-learning to learn to perform the perceptual matching task (Sutton & Barto, 1998; Watkins, 1989). Without going into the mathematical details, the reader should note that Q-learning is an approximate dynamic programming method (Si, Barto, Powell, & Wunsch, 2004). It is easy to show that, under mild conditions, the sequence of actions found by an agent using Q-learning is guaranteed to converge to an optimal sequence found by dynamic programming (Watkins & Dayan, 1992). Hence, the benchmarks based on dynamic programming and on Q-learning are related.

2.2.4. Reinforcement learning and task performances

In a reinforcement-learning framework, it is assumed that an agent attempts to choose actions so as to receive the most reward possible. The agent explores its environment by assessing its current state and choosing an action. After executing this action, the agent will

be in a new state and will receive a reward (possibly zero) associated with this new state. The agent adapts its behavior in a trial-by-trial manner by noticing which actions tend to be followed by future rewards and which actions are not.

To choose good actions, the agent needs to estimate the long-term reward values of selecting possible actions from possible states. Ideally, the value of selecting action a_t in state s_t at time t, denoted $Q(s_t, a_t)$, should equal the sum of rewards that the agent can expect to receive in the future if it takes action a_t in state s_t:

$$Q(s_t, a_t) = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\right]$$

where t is the current time step, k is an index over future time steps, r_{t+k+1} is the reward received at time $t + k + 1$, and γ ($0 < \gamma \leq 1$) is a term that serves to discount rewards that occur in the far future more than rewards that occur in the near future. An agent can learn accurate estimates of these ideal values on the basis of experience if it updates its estimates at each time step using the Q-learning update equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]$$

where the agent makes action a_t in state s_t and receives reward r_{t+1}, and α is a step size or learning rate parameter (Sutton & Barto, 1998; Watkins, 1989).

In our first set of simulations, a reinforcement-learning agent was trained to perform the perceptual matching task as follows. At each time step, the state of the agent represented the difference in shape between the comparison and target objects. The state was a three-dimensional vector whose elements were set to the values of the shape parameters for the comparison object minus the values of these parameters for the target object. Six possible actions were available to the agent corresponding to the six buttons that a subject could press to modify the action variables. The agent chose an action using an ε-greedy strategy, meaning that the agent chose the action a that maximized $Q(s_t, a)$ with probability $1 - \varepsilon$ (ties were broken at random), and chose a random action with probability ε. The value of ε was initialized to one, and then it was slowly decreased during the course of a simulation. (Specifically, ε was set to $1/N$ where N is the action number. That is, ε was set to 1 at the first action, $1/2$ at the second action, $1/3$ at the third action, and so on.) As a result, the agent tended to often ''explore'' a wide range of actions toward the beginning of a simulation, and tended to ''exploit'' its current estimates of the best action to take toward the middle and end of a simulation. If the agent chose an action that caused the comparison object to have the same shape as the target object, the agent received a reward of 100. Otherwise, it received a reward of -1. The agent performed the training trials of the experiment in the same manner as our human subjects—it performed seven blocks of training trials with 26 trials per block. At the start of each simulation, its ''Q-values'' were initialized to zero, its discount rate γ was set to 0.7, and its learning rate α was set to 0.45. In preliminary simulations, these values were found to be best in the sense that they led to performances that most closely matched human performances (i.e., they led to performances that minimized the

sum of squared differences between the average number of steps used by the agent at each block and the average number of steps used by subjects). To accurately estimate the agent's performances during training, the agent was simulated 1000 times.

For brevity, only the results for the training trials using the two linear causal structures are reported here (similar results were obtained for the other causal structures). The leftmost graph in Fig. 9 shows the learning curves of the reinforcement-learning agent (referred to as the model-free agent; see blue line) and of the human subjects that participated in our experiment (gray line). The horizontal axis of this graph plots the block number, and the vertical axis plots the average difference between the number of steps (i.e., actions or button presses) used by the agent or by human subjects during a trial and the optimal number of steps for that trial as computed by the dynamic programming procedure (as in Fig. 5). Interestingly, the learning curves of the simulated agent and of the human subjects have similar shapes, though subjects outperformed the agent at nearly all stages of training. Modifications of the agent by either using different values for the agent's parameters or by adding ''eligibility traces''[4] did not significantly alter this basic finding.

The rightmost graph of Fig. 9 shows the average reversed-order rates as a function of the block number for the agent and the subjects. Whereas subjects' rates declined during the course of training, suggesting an increase in their causal knowledge, the agent's rates remained relatively constant.

Why did human subjects show better learning performances than the simulated agent? In the Artificial Intelligence literature, a distinction is made between model-free versus model-based reinforcement-learning agents. The agent described above is an instance of a

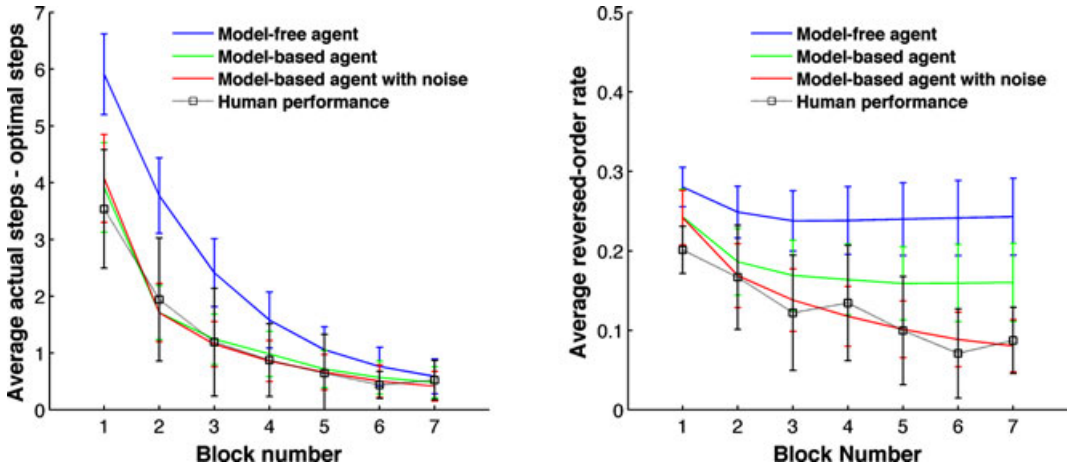

Fig. 9. (Left) Learning curves for the model-free agent (blue line), the model-based agent (green line), the model-based agent with noise (red line), and for human subjects (gray line) on the training trials for experimental conditions with linear causal structures. The horizontal axis plots the block number, and the vertical axis shows the average difference between the number of steps (i.e., button presses) used by an agent or by subjects during a trial and the optimal number of steps for that trial as computed by the dynamic programming procedure. (Right) Average reversed-order rates for the reinforcement-learning agents and for human subjects.

model-free agent. Although model-free agents are more common in the literature, we hypothesized that a model-based reinforcement-learning agent may provide a better account of our subjects' performances. Previous researchers have hypothesized that people are capable of both model-free and model-based reinforcement learning (Daw, Niv, & Dayan, 2005; Gläscher, Daw, Dayan, & O'Doherty, 2010). Model-based agents typically learn faster than model-free agents, albeit with greater computational expense. Based on real-world experiences, a model-based agent learns an internal model of how its actions influence the environment. Importantly, the agent updates its Q-values from both real-world experiences with the environment and from simulated experiences with the model (see Sutton & Barto, 1998, for details).

In our next set of simulations, we implemented a model-based reinforcement-learning agent. The agent's model was an artificial neural network which learned a mapping from actions to changes in shape parameters. Its six input units corresponded to the six possible actions or key presses (an action variable could either increase or decrease in value, and there were three action variables). Its nine output units corresponded to the nine possible influences on the comparison objects' shape parameters (a shape parameter could either increase in value, decrease in value, or maintain the same value, and there were three shape parameters). The three output units corresponding to each shape parameter used the softmax activation function, meaning that their activations provided a multinomial distribution over the possible influences on the parameter. The network did not contain any hidden units.

When updating its Q-values, the model-based agent used ''prioritized sweeping'' (Moore & Atkeson, 1993). This is an efficient method for focusing Q-value updates to state-action pairs associated with large changes in expected reward. Large changes occur, for example, when the current state is a non-goal state and the agent discovers a previously unfamiliar action that leads to a goal state. Large changes also occur when the current state is a non-goal state, and the agent discovers a new action that leads to a new non-goal state known to lie on a path toward a goal state.

The prioritized sweeping algorithm is described in Fig. 10. In brief, our simulations used prioritized sweeping as follows. At each moment in time, the model-based agent maintained a queue of state-action pairs whose Q-values would change based on either real or simulated experiences. For each update based on a real experience, there were up to N updates based on simulated experiences. The items on the queue were prioritized by the absolute amount that their Q-values would be modified. Suppose that at some moment in time, state-action pair (s^*, a^*) had the highest priority. Then $Q(s^*, a^*)$ would be updated. If performing this update on the basis of simulated experience, the agent used the model to predict the resulting new state. In addition, the agent also used the model to examine changes to the Q-values for all state-action pairs predicted to lead to state s^*, known as predecessor state-action pairs. These predecessor state-action pairs were added to the queue, along with their corresponding priorities.

The simulations with the model-based agent were conducted in an identical manner to those with the model-free agent. However, the model-based agent used different parameter values. Its discount rate γ was set to 0.3, its learning rate α was set to 0.05, and N, the

Loop forever:
(a) $s \leftarrow$ current (non-goal) state
(b) $a \leftarrow$ action chosen via ε-greedy strategy
(c) Execute action a, observe resultant state s' and reward r
(d) Update Model based on action a and state s'
(e) $p \leftarrow |r + \gamma \max_{a'} Q(s', a') - Q(s, a)|$
(f) Insert s, a into PQueue with priority p
(g) Repeat N times, while PQueue is not empty:
  a. $s, a \leftarrow$ first(PQueue)
  b. $s' \leftarrow$ Model(s, a)
  c. $r \leftarrow$ 100 if s' is goal state, -1 otherwise
  d. $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
  e. Repeat for all $\bar{s}$, $\bar{a}$ predicted by Model to lead to s:
    i. $\bar{r} \leftarrow$ 100 if $\bar{s}$ is goal state, -1 otherwise
    ii. $p \leftarrow |\bar{r} + \gamma \max_{a} Q(s, a) - Q(\bar{s}, \bar{a})|$
    iii. Insert $\bar{s}$, $\bar{a}$ into PQueue with priority p

Fig. 10. Prioritized sweeping algorithm (adapted from Sutton & Barto, 1998).

number of Q-value updates based on simulated experiences for each update based on a real experience, was set to 5. In preliminary simulations, these values were found to be best in the sense that they led to performances that most closely matched human performances.

The results for the two experimental conditions using linear causal structures are shown in the leftmost graph of Fig. 9 (once again, results for the other conditions were similar). The learning curves for the model-based agent (green line) and for human subjects (gray line) are nearly identical. Because the model-based agent provides a better account of subjects' performances than the model-free agent discussed above (based upon the sum of squared differences between the average number of steps used by an agent at each block and the average number used by subjects), the results suggest that our subjects may have also built and used internal models of how their actions influenced the environment. The rightmost graph of Fig. 9 shows the average reversed-order rates. Although the model-based agent performed as well as people on the experimental task, its reversed-order rate was significantly greater than that of human subjects.

A possible explanation for the model-based agent's large reversed-order rate was hinted at above.[4] If an agent has a perfect model of how its actions influence the environment, then it may be possible for the agent to perform well using button presses in which a subordinate action variable is modified before a superordinate variable. However, if there is uncertainty in the model, then better performance will be achieved if superordinate variables are changed first. Thus, it may be that the model-based agent had larger reversed-order rates than subjects because the agent learned a perfect model (or nearly so), whereas subjects' models were imperfect or uncertain.

To test this hypothesis, we simulated a third reinforcement-learning agent, referred to as the model-based agent with noise. This agent was identical to the original model-based agent, except that noise was added when training its neural network model of how actions influence the environment. The three output units corresponding to a shape parameter received the correct target signals with probability p (we set $p = 0.95$ in our simulations).

With probability $1 - p$, the units received target signals indicating that the shape parameter did not change its value. This type of noise was intended to mimic a situation in which the agent noticed perceptual changes to a comparison object corresponding to a change in the value of a shape parameter with probability p. However, the agent failed to notice these changes with probability $1 - p$. When the agent failed to notice the changes, this could be referred to as a lapse of attention.

The results for the two conditions using linear causal structures are shown in the leftmost graph of Fig. 9. The learning curves for the model-based agent with noise (red line) and for human subjects (gray line) are nearly identical. The rightmost graph shows the average reversed-order rates. The rates of the model-based agent with noise and of subjects are also nearly identical. On the basis of these data, we conclude that the model-based agent with noise provides a good account of subjects' performances on the perceptual matching task.

3. Conclusions

Sequential action tasks are commonplace in our everyday lives. In this article, we studied whether people were successful at learning to perform a perceptual matching task, an instance of a sequential action task. We used two benchmarks to evaluate the quality of subjects' learning. One benchmark was based on optimal performance as defined by a dynamic programming procedure. The other was based on an adaptive computational agent that used reinforcement learning to learn to perform the task. Overall, our analyses suggest that subjects learned to perform the perceptual matching task in a near-optimal manner. When doing so, subjects learned, at least partially, the causal structure underlying the task. In addition, subjects' performances were broadly consistent with those of model-based reinforcement-learning agents. These agents learned internal models of how their actions influenced the external environment, and they used these models to reason about good actions to take at each moment in time.

From a methodological perspective, the research reported here is notable for the way it combines experimental and computational studies of people's performances on a sequential action task. Comparisons of people's performances with either optimal performance as calculated by dynamic programming or with learning performances of reinforcement-learning agents are relatively unusual in the scientific literature. We believe that this article highlights the advantages of conducting both types of comparisons. We hope that other researchers will also include both types of comparisons in their future studies of people's performances on sequential action tasks.

Conceptually, the research reported here is also notable. In the Artificial Intelligence literature, model-free reinforcement-learning agents are significantly more common than model-based agents. Here, we found that model-based agents provided a better account of our experimental data. This result suggests that our subjects did not learn to perform the perceptual matching task by simply correlating experimental states and actions with predictions of future reward. Instead, subjects learned detailed models of the influences of actions on their environments, or of how actions altered one experimental state into another state. Moreover,

subjects were able to use these models to reason about good actions to take at each moment in time.

Above we argued that subjects achieved near-optimal performance on the experimental task because they learned the causal structure underlying this task. If so, then it is worth probing further into precisely what they learned. Did they learn causal relations among the shape parameters, such as X has a causal influence on Y which, in turn, has a causal influence on Z? Or did they learn causal relations between action variables and shape parameters, such as action variable A has a causal influence on shape parameters X, Y, and Z? Importantly, these two possibilities are exact notational variants of each other using the notation of Bayesian networks (i.e., the Bayesian networks corresponding to these two possibilities characterize the same joint probability distribution over all variables). Based on the experimental data collected here, we cannot distinguish which possibility better describes what our subjects learned. We have consistently described the causal relations in the experimental task in terms of the former possibility (causal relations among shape parameters) for ease of explanation.

We are also unable to distinguish the type of mental representation that our subjects used to represent these causal relations. Did subjects represent their causal knowledge as Bayesian networks, sets of logical rules, or sets of associations? Again, our experimental data does not allow us to answer this question.

Moreover, we are unable to conclude that subjects found some causal structures easier to learn than others. Although this possibility seems plausible, statistical tests of this possibility did not reach our threshold for statistical significance. Future work will need to address this issue.

The cognitive science literature now contains several studies of human performances on sequential action tasks. Some studies have suggested that human performances are optimal, whereas other studies have suggested the opposite. To date, the field of Cognitive Science does not have a good understanding of the factors influencing whether people will achieve optimal performance on a sequential action task. Future research will need to focus on this critical issue. In the Introduction section of this article, we mentioned that perceptual aliasing (Stankiewicz et al., 2006) or the existence of actions leading to large rewards in the short-term but not the long-term (Gureckis & Love, 2009b; Neth et al., 2005) seem to be factors leading to suboptimal performances. Here, we propose a new understanding of when people will (or will not) achieve optimal performances on sequential action tasks. We hypothesize that people will achieve near-optimal performances on sequential action tasks—especially sensorimotor tasks with underlying latent causal structures—when they can detect the effects of their actions on the environment, and when they can represent and reason about these effects using an internal mental model.

Future research will need to test this hypothesis by considering extreme instances of sequential action tasks. For example, how will people perform when long sequences of actions are needed to achieve a goal? How will they perform when there are long temporal delays between the execution of an action and the effect of that action on the environment? Or how will they perform when the effects of actions are only partially observable? Experiments addressing these questions may highlight the importance of developing new teaching

or training procedures. If people perform suboptimally on a sequential action task, it may be possible to develop new training procedures that enable people to boost their performances so they are closer to optimal. If so, we conjecture that these procedures will be effective because they better allow people to detect and understand the causal effects of their actions on the environment.

Notes

1. Trials on which a subject failed to convert the comparison shape to the target shape within 10 button presses were ignored in the analyses discussed below. Such trials were more common in early stages of training (about 7 times in Block 1), and relatively uncommon in later stages of training (about 0–1 times in Blocks 4–7).
2. This result motivates our use of deterministic causal influences among variables. Stochastic influences would have made a difficult task even harder to perform.
3. A perhaps subtle point is that if a subject has perfect knowledge of the causal relations among the shape parameters, then it may be possible for the subject to perform optimally using button presses in which a subordinate action variable is modified before a superordinate variable. But if there is uncertainty in a subject's causal knowledge, as was the case with our subjects, then better performance will be achieved (on average) if superordinate variables are modified first.
4. Eligibility traces allow an agent to remember which actions it has selected in the recent past, and to use this information to help credit actions which lead to reward.

Acknowledgments

We thank the anonymous reviewers and the editor for helpful comments on this manuscript, especially with respect to the reinforcement-learning simulations. This work was supported by a Grant-in-Aid for Scientific Research (20730480) from the Japan Society for the Promotion of Science, by a research grant from the Air Force Office of Scientific Research (FA9550-06-1-0492), and by a research grant from the National Science Foundation (DRL-0817250).

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*, 723–767.

Bellman, R. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Busemeyer, J. R. (2001). Dynamic decision making. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 3903–3908). Oxford, UK: Pergamon.

Busemeyer, J. R., & Pleskac, T. J. (2009). Theoretical tools for understanding and aiding dynamic decision making. *Journal of Mathematical Psychology*, *53*, 126–138.

Chhabra, M., & Jacobs, R. A. (2006). Near-optimal human adaptive control across different noise environments. *The Journal of Neuroscience*, *26*, 10883–10887.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to algorithms* (2nd ed.). Cambridge, MA: MIT Press.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.

Daw, N. D., & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, *14*, 2567–2583.

Fu, W.-T., & Anderson, J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, *135*, 184–206.

Geisler, W. S. (2004). Ideal observer analysis. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 825–837). Cambridge, MA: MIT Press.

Gibson, F. P., Fichman, M., & Plaut, D. C. (1997). Learning in dynamic decision tasks: Computational model and empirical evidence. *Organizational Behavior and Human Decision Processes*, *71*, 1–35.

Gielis, J. (2003). A generic geometric transformation that unifies a wide range of natural and abstract shapes. *American Journal of Botany*, *90*, 333–338.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*, 585–595.

Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior*, *21*, 273–286.

Gopnik, A., & Shulz, L. (2007). *Causal learning: Psychology, philosophy, and computation*. New York: Oxford University Press.

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*, 461–482.

Gureckis, T. M., & Love, B. C. (2009a). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, *53*, 180–193.

Gureckis, T. M., & Love, B. C. (2009b). Short-term gains, long-term pains: Reinforcement learning in dynamic environments. *Cognition*, *113*, 293–313.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 1–26.

McCallum, R. A. (1993). Overcoming incomplete perception with utile distinction memory. In *Proceedings of the 10th International Machine Learning Conference*, pp. 190–196. San Francisco: Morgan Kaufmann.

Moore, A., & Atkeson, C. (1993). Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, *13*, 103–130.

Neth, H., Sims, C. R., & Gray, W. D. (2005). Melioration despite more information: The role of feedback frequency in stable suboptimal performance. In *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society*, pp. 627–632. Santa Monica, CA: Human Factors and Ergonomics Society.

Neth, H., Sims, C. R., & Gray, W. D. (2006). Melioration dominates maximization: Stable suboptimal performance despite global feedback. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, pp. 627–632. Mahwah, NJ: Erlbaum.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1598.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of melioration and rational choice. *Journal of Behavioral Decision Making*, *15*, 233–250.

Si, J., Barto, A. G., Powell, W. B., & Wunsch, D. (2004). *Handbook of learning and approximate dynamic programming*. Piscataway, NJ: Wiley-IEEE.

Stankiewicz, B. J., Legge, G. E., Mansfield, J. S., & Schlicht, E. J. (2006). Lost in virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 688–704.

Stanley, W. B., Mathews, R. C., Buss, R. R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction, and practice in a simulated process control task. *Quarterly Journal of Experimental Psychology*, *41A*, 553–577.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, *112*, 159–192.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*, 907–915.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. Unpublished doctoral dissertation. Cambridge, UK: Cambridge, University.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292.

Whitehead, S. D., & Ballard, D. H. (1991). Learning to perceive and act by trial and error. *Machine Learning*, *7*, 45–83.