Psychological Review
2006, Vol. 113, No. 2, 390–408

0033-295X/06/$12.00 DOI: 10.1037/0033-295X.113.2.390

Using Reinforcement Learning to Understand the Emergence of "Intelligent" Eye-Movement Behavior During Reading

Erik D. Reichle and Patryk A. Laurent
University of Pittsburgh

The eye movements of skilled readers are typically very regular (K. Rayner, 1998). This regularity may arise as a result of the perceptual, cognitive, and motor limitations of the reader (e.g., limited visual acuity) and the inherent constraints of the task (e.g., identifying the words in their correct order). To examine this hypothesis, reinforcement learning was used to allow an artificial "agent" to learn to move its eyes to read as efficiently as possible. The resulting patterns of simulated eye movements resembled those of skilled readers and suggest that important aspects of eye-movement behavior might emerge as a consequence of satisfying the constraints that are imposed on readers. These results also suggest novel interpretations of some contentious empirical results, such as the fixation duration costs associated with word skipping (R. Kliegl & R. Engbert, 2005), and theoretical assumptions, for example the familiarity check in the E-Z Reader model of eye-movement control (E. D. Reichle, A. Pollatsek, D. L. Fisher, & K. Rayner, 1998).

Keywords: attention, eye-movement control, computational models, reading, reinforcement learning

How do readers learn to coordinate the perceptual, cognitive, and motor processes that control eye movements during reading? One might wonder why this is an interesting question given that moving one's eyes during reading seems effortless and, in contrast to most other reading skills (e.g., decoding), is not explicitly taught. Indeed, most people rarely think about how they move their eyes during reading and, if queried, often report that their eyes simply move smoothly and continuously across each line of text, with occasional pauses or movements backward to places in the text that are especially interesting or difficult to understand. Most people would therefore find it surprising to learn that more than two decades of research have demonstrated that most of these intuitions regarding eye movements during reading are erroneous (for an extensive review of this research, see Rayner, 1998). Evidence instead shows that the perceptual, cognitive, and motor processes that guide the eyes during reading are remarkably complex, and that these processes must be carefully coordinated. What makes this even more remarkable is that we are not biologically programmed to read or to move our eyes so as to make reading possible but must instead learn to do so through years of practice. Knowing these facts, one gains an appreciation for how interesting our question really is. Our goal in this article, therefore, is to provide one account for how the various perceptual, cognitive, and motor processes that control the eyes become coordinated to allow skilled reading.

As we have asserted, eye-movement behavior during reading is surprisingly complex (Rayner, 1998; Rayner & Pollatsek, 1989). For example, consider the simple fact that as you read this sentence, your eyes make a series of very rapid ballistic movements (called *saccades*), each lasting 20–35 ms, and pause briefly once or twice on most of the words in the sentence (Rayner, 1978, 1998). These pauses (called *fixations*) provide brief intervals during which information can be extracted from the page. Although most fixation durations are 200–300 ms, fixations as short as 50 ms do occur as do fixations as long as 500 ms (Rayner, 1978, 1998). Because useful visual information cannot be extracted from the page during the actual saccades (Ishida & Ikeda, 1989; Wolverton & Zola, 1983), reading is like a slide show in which each slide (i.e., the information available from a given viewing location) is visible for about a quarter of a second. Although most saccades move the eyes forward approximately 7–9 character spaces, approximately 15% of the saccades are *regressions*, which move the eyes back to earlier parts of the text. Regressions are thought to stem from both problems with higher level linguistic processing, such as when readers make incorrect syntactic analyses (Frazier & Rayner, 1982), and incomplete word identification (Engbert, Longtin, & Kliegl, 2002). Finally, because saccades are prone to motor error, the distribution of fixation locations tends to resemble truncated Gaussians that are centered near the middles of words, with the missing tails being due to those cases where the saccades

Erik D. Reichle, Department of Psychology, Learning Research and Development Center, and Center for the Neural Basis of Cognition, University of Pittsburgh; Patryk A. Laurent, Center for Neuroscience and Center for the Neural Basis of Cognition, University of Pittsburgh.

Portions of this work were presented at Architectures and Mechanisms for Language Processing, 2004, Aix-en-Provence, France. This work was supported by the Department of Education, Institute of Education Sciences Grant R305G020006, and by Short Development Grant IBN040003P from the Pittsburgh Super Computing Center. Java source code to run the reading agent is available at http://rl.pakl.net

We would like to thank Remi Coulom for answering our questions about reinforcement learning algorithms and Jane Ashby, Ralf Engbert, Ibrahim Hakki, Reinhold Kliegl, Charles Perfetti, Alexander Pollatsek, Keith Rayner, and especially Tessa Warren for their helpful comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Erik D. Reichle, Department of Psychology, University of Pittsburgh, 635 LRDC, 3939 O'Hara Street, Pittsburgh, PA 15260. E-mail: reichle@pitt.edu

presumably either under- or overshot their intended target words (McConkie, Kerr, Reddix, & Zola, 1988; McConkie, Zola, Grimes, Kerr, Bryant, & Wolff, 1991; Nuthmann, Engbert, & Kliegl, 2005; Rayner, 1979; Rayner, Sereno, & Raney, 1996).

Although the above description of eye-movement behavior during reading indicates that it is highly complex (for a discussion of whether this complexity reflects inherent stochasticity or deterministic nonlinearities of the oculomotor and/or lexical processing systems, see Engbert, Kliegl, & Longtin, 2004), it is also remarkably systematic. We are interested in how this regularity develops. The hypothesis that we test in this article is that this regularity develops as the perceptual, cognitive, and motor processes that mediate text comprehension during reading become increasingly coordinated. If this is correct, then it suggests that the regularity that is observed in the eye movements of a skilled reader may come about as the result of attempting to satisfy two goals—to read as quickly as possible while at the same time maintaining some level of comprehension. To explore this hypothesis, it is important to first consider the factors that constrain a reader in achieving these two goals.

The first constraint is that the high visual acuity that is required for the identification of printed words is largely limited to a small region of the retina—the *fovea*, which extends across the central 2° of the visual field. Although words can be identified outside of this region, word identification becomes slower and less accurate as the angular distance between the word and the fovea increases (Lee, Legge, & Ortiz, 2003; Rayner & Bertera, 1979; Rayner & Morrison, 1981). This provides an answer to why most words are fixated during reading (Rayner, 1978, 1998): Doing so ensures that the lexical processing that is necessary to identify a word will be completed from the fovea. It also suggests that at least some of the regularity that is observed in the eye movements of readers may reflect the fact that, in order to understand whatever is being read, it is first necessary to identify each of the words on the page (or at least most of them).

The second factor that limits reading speed is the amount of time required to identify words. Evidence indicates that there is a considerable amount of variability in the time required to identify words (e.g., 150–300 ms; Rayner & Pollatsek, 1989; Sereno, Rayner, & Posner, 1998). This variability stems from between-individual differences in reading ability (Perfetti, 1985; Perfetti & Hart, 2002) as well as other variables that affect the difficulty of word processing. For example, words that are frequently encountered in text may be represented better in memory and hence are easier to identify. This hypothesis is consistent with numerous eye-tracking experiments showing that more common words tend to be the recipients of fewer and shorter fixations than are less common words (Altarriba, Kroll, Sholl, & Rayner, 1996; Henderson & Ferreira, 1990; Hyönä & Olson, 1995; Inhoff & Rayner, 1986; Just & Carpenter, 1980; Kennison & Clifton, 1995; Kliegl, Grabner, Rolfs, & Engbert, 2004; Rayner, 1977; Rayner, Ashby, Pollatsek, & Reichle, 2004; Rayner & Duffy, 1986; Rayner & Raney, 1996; Schilling, Rayner, & Chumbley, 1998; Sereno, 1992). Other variables that have also been shown to affect fixation durations and probabilities include word length (Brysbaert & Vitu, 1998; Just & Carpenter, 1980; Kliegl et al., 2004; Rayner & McConkie, 1976; Rayner et al., 1996), the age at which a given word is learned (Juhasz & Rayner, 2003, in press), and the predictability of a given word within its sentence context (Balota, Pollatsek, & Rayner, 1985; Ehrlich & Rayner, 1981; Kliegl et al., 2004; Rayner et al., 2004; Rayner & Well, 1996). Other things being equal, words that are shorter in length, learned earlier in life, or predictable from their linguistic context tend to receive fewer and shorter fixations than words that are longer, learned later in life, or less predictable.

Of course, the claim that fixation durations and probabilities reflect how long it takes to identify the individual words during reading is predicated on the assumption that the process of identifying words is linked to the moment-to-moment "decisions" about when to move the eyes from one word to another;[1] otherwise, one would not expect any systematic relationships between those variables that presumably affect word-processing difficulty, on one hand, and fixation durations and probabilities, on the other. As it turns out, there is some empirical evidence for this eye-mind assumption: Schilling et al. (1998) found that in natural reading, high-frequency words were the recipient of fewer and shorter fixations than low-frequency words, and that the same high-frequency words also required less time to pronounce and make lexical decisions about than the low-frequency words. Thus, at least the normative frequency with which a word occurs in printed text seems to modulate word-processing difficulty and how long a word will be looked at during reading—a finding that is consistent with the existence of an eye-mind link.

There is also theoretical precedent for assuming an eye-mind link. Many of the computational models that have been successful in explaining eye movement control during reading assume that some aspect of lexical processing either directly or indirectly affects the moment-to-moment decisions about when to move the eyes (Engbert et al., 2002; Engbert, Nuthmann, Richter, & Kliegl, in press; Just & Carpenter, 1980, 1987; McDonald, Carpenter, & Shillcock, in press; Pollatsek, Reichle, & Rayner, in press; Reichle et al., 1998; Reichle, Rayner, & Pollatsek, 1999, 2003, 2006; Reilly, 1993; Reilly & Radach, 2003; Salvucci, 2000; Thibadeau, Just, & Carpenter, 1982). (For a comparative review of these and other models of eye movement control during reading, see Reichle et al., 2003.) Although these models have been extremely useful for both explaining existing data and guiding new research (Rayner, Pollatsek, & Reichle, 2003; see also Reichle et al., 2006), all of the models are based on controversial a priori assumptions. The models are also limited in that they describe the behavior of *skilled* readers; that is, the models specify how various perceptual, cognitive, and motor processes account for the eye-movement behavior of readers who have had many years of reading experience. These models do not explain how eye-movement behavior comes to exhibit the regularity that is observed with skilled readers.

In this article, we attempt to overcome these limitations and provide an account of how systematic eye movements might

[1] Our use of term *decisions* is not meant to imply that readers are consciously deciding when and where to move their eyes during reading. Skilled readers instead program and execute saccades with little (if any) conscious reflection or effort. This is probably due to the fact that as readers become more skilled at reading, lexical processing and eye-movement control become increasingly coordinated so that, after years of practice, the "decisions" about when and where to move the eyes become largely automatic.

emerge as an adaptive solution to the problem of learning how to read efficiently. We do this by presenting the results of two simulations that were completed using an "agent" or artificial "reader" that is capable of learning to control its eye-movement behavior so as to read efficiently. The agent was developed around a minimum number of a priori assumptions that were intended to reflect constraints that can be motivated by independent physiological and psychological evidence. In contrast to models like E-Z Reader (Pollatsek et al., in press; Rayner et al., 2004; Reichle et al., 1998, 2003, 2006) and SWIFT (Engbert et al., 2002, in press), our starting assumptions were not selected so at to guarantee that the agent's eye-movement behavior would resemble that of actual readers; our approach is instead similar to that used by Legge, Klitz, and Tjan (1997; Legge, Hooven, Klitz, Mansfield, & Tjan, 2002; Klitz, Legge, & Tjan, 2000) in their simulations using an ideal observer (i.e., their Mr. Chips model). However, whereas those ideal-observer simulations attempted to explain how lexical knowledge, visual acuity limitations, and oculomotor constraints affect *where* the eyes move during reading, our simulations also attempt to explain how these same variables affect *when* the eyes move during reading. In the next section, we describe the reading agent and our starting assumptions; the simulation results that were obtained using the agent are then described in the subsequent sections.

Reinforcement Learning and the Adaptive Reading Agent

The simulations that are reported in this article were intended to determine if an adaptive reading agent that is capable of learning how to control its own eye movements could do so in a way that leads to efficient reading. The adaptive reading agent is a computational system that we hypothesized should be capable of learning how to control both when and where it moves its eyes. Our operational definition of "efficient reading" was that all of the words in a given sentence had to be identified as quickly as possible, in their serial order. We assumed that these are the minimal requirements to ensure comprehension.

It is also important to understand what the present simulations were intended to demonstrate: We made no attempt to provide quantitative fits to any data but instead focused our efforts on showing how a few simple principles can give rise to the qualitative patterns of eye movements that are observed with skilled readers.[2] The agent should not be conceptualized as an alternative model of eye-movement control but should instead be thought of as a way of demonstrating how some of the more important properties of these models might develop through years of reading experience. (We will say more about this point in the General Discussion section.) The simulations were meant to be an existence proof demonstrating that a small set of constraints are sufficient to produce patterns of eyes movements that share many of the characteristics of the eye movements that are observed with real human readers. The constraints that were imposed upon the agent were as follows:

Saccadic Eye Movements

The first pair of constraints concerned saccades: They require a certain amount of time to program, and only one saccade can be programmed at a time. For simplicity, we assumed that saccades take three time steps to program and one time step to execute. Although these values are arbitrary, they are consistent with the fact that saccades take an appreciable amount of time to program (125–150 ms; Becker & Jürgens, 1979; Rayner, Slowiaczek, Clifton, & Bertera, 1983) but less time to execute (approximately 25–35 ms). The assumption that only a single saccade can be programmed at a time was adopted to keep the simulations as simple as possible.[3] For the same reason, the first simulation did not include saccadic error; we instead assumed that saccades always move the eyes to the intended fixation location. (We explored the effects of introducing saccadic error in our second simulation.) The agent was allowed to program saccades of up to 10 character spaces to either the left or the right of its current fixation location. Although the values of this saccade-length restriction were—again—arbitrary, some fixed values were necessary to run the simulation, and a maximum saccade length of 10 was adequate to allow the agent to skip words if such an action was optimal. Finally, lexical processing was not allowed to continue during saccade execution. This last assumption was adopted to simplify the modeling and may or may not be consistent with what actually happens with human readers; however, the effects of this assumption on the agent's behavior were expected to be negligible (for a discussion of this issue, see Pollatsek et al., 2005).

Word Identification

The second set of constraints was related to word identification: Each word requires a certain amount of time to be identified, and this time is modulated by lexical variables and visual acuity. Because we wanted to determine if our agent would learn to anticipate how long it takes to identify words and then use this information in deciding when and where to move its eyes, it was important that our simulations include words that varied in terms of their processing difficulty. Of course, a real reader who is learning to anticipate how long it will take to identify words faces a considerable challenge because processing difficulty is related to many variables (e.g., word frequency, age-of-acquisition, and so forth) and because these variables tend to be intercorrelated (e.g., high-frequency words tend to be shorter than low-frequency

[2] Because our simulations only provided a qualitative account of eye-movement data, the simulations were completed using arbitrary time units for measuring both process durations (i.e., the time needed to program saccades and identify words) and fixation durations.

[3] There is a considerable amount of evidence from behavioral experiments that saccades are actually programmed in two stages—a *labile* stage that can be canceled by a subsequent saccade program, and a *nonlabile* stage that is not subject to cancellation (Becker & Jürgens, 1979; Leff, Scott, Rothwell, & Wise, 2001; McPeek, Skavenski, & Nakayama, 2000; Molker & Fischer, 1999). The nonlabile stage can be thought of as a "point of no return," or the point in time when a program to move the eyes from one location to another has become obligatory and will be executed irrespective of whether another program to move the eyes to a second location has been initiated. Because we wanted to keep our simulations as simple as possible, we did not implement both stages of saccadic programming; we instead simply assumed that upon being initiated, saccades could not be canceled and are thus obligatory. However, it is important to note that our assumption that only one saccade can be programmed at a time is inconsistent with evidence suggesting that two saccades can sometimes be programmed in parallel (Vergilino & Beauvillain, 2000).

words). However, to keep our simulations as simple as possible, we made each word's processing difficulty related to its length, and not its frequency, age-of-acquisition, or predictability.[4] Our simulation "sentences" were thus constructed so that they included words that varied along four levels of difficulty: 1-, 3-, 5-, and 7-letter words were assumed to require four, five, six, and seven time steps to identify, respectively. (These numbers for word-identification times excluded the effects of visual acuity, which are discussed next.) Our decision to make processing difficulty a function of length also allowed a more sensitive test of the agent's capacity to learn how to anticipate word-processing difficulty because this variable is perfectly predicted by word length (i.e., the correlation between word length and processing difficulty is $r = +1$). Although our manipulation of word length could be construed as being a purely visual one and by implication have little to say about the relationship between cognition and eye-movement control, it is important to note that the amount of time that is actually needed to identify a given word could in principle be a function of visual or cognitive variables or both (e.g., frequency of occurrence), and what is important is that there are reliable cues that the agent can use to learn how long it will take to identify a given word. It is therefore important that the time required to identify words—irrespective of whether it is a function of perceptual or cognitive (or both) variables—be related to some perceivable variable because the agent has no a priori knowledge about how long it will take to identify individual words but must instead learn that it takes its word identification system more time to identify some (e.g., longer) words than other (e.g., shorter) words and then adapt its eye-movement behavior to take advantage of this fact.

Visual Acuity

Another constraint concerned visual acuity: The high visual acuity that is needed to rapidly identify words is spatially limited (Lee et al., 2003; Rayner & Bertera, 1979; Rayner & Morrison, 1981). This assumption is consistent with results showing that words that are briefly displayed in isolation can be identified most rapidly if they are presented so that the word is fixated on its center (O'Regan & Lévy-Schoen, 1987; O'Regan, Lévy-Schoen, Pynte, & Brugaillière, 1984). These results suggest that readers direct their eyes toward the centers of words because this is the *optimal viewing position*, or position from where words can be identified most rapidly (O'Regan, 1990, 1992; O'Regan & Lévy-Schoen, 1987), although the means of the fixation landing-site distributions tend to be closer to the beginnings of words, so that the *preferred viewing location* in natural reading is to the left of the optimal viewing position (McConkie et al., 1988, 1991; Rayner, 1979; Rayner et al., 1996). To keep our simulations simple, we assumed that there is no additional processing cost in cases where the centers of words are fixated, and that there is a linear relationship between visual acuity and distance, with the time that is needed to identify a given word increasing by one time step for each character space that the eyes deviate from the center of the word that is being processed. The rate of lexical processing is thus attenuated by an amount that is directly proportional to the distance between the fixation location and the center of the word that is being identified (Lee et al., 2003; Rayner & Bertera, 1979; Rayner & Morrison, 1981).

In the event of a saccade, the time that has been spent identifying the current word is discounted or prorated according to the change in visual acuity that resulted from the saccade. This ensures that time spent identifying a word from an eccentric position is not worth as much as time spent identifying a word from the optimal viewing position. For example, imagine that the agent starts processing a seven-letter word (which takes seven time steps to identify from the optimal viewing position) from the first letter of this word. Now imagine that, upon fixating the word, the agent immediately requests a saccade to move its eyes to the center of the word. Because the initial viewing location is three character spaces from the optimal viewing position, it would ordinarily take 10 time steps to identify the word from the initial viewing location. The amount of time that was actually spent processing the word from this location (i.e., the three time steps needed to program the refixation saccade) is thus equal to 3/10 of the total time needed to identify the word. This proportion of the total processing time required to identify the word and that has been completed from the first viewing location (rounded down to the nearest integer value; i.e., $3/10 \times 7 = 2.1$, which is rounded down to 2) is then subtracted from the total time that is needed to identify the word (seven time steps) when the eyes land upon the second viewing location. The amount of processing done from the first viewing location is thus prorated to reflect the slowing of processing that results from poor visual acuity. The proportion of the total processing that is completed is always rounded down to be conservative; this safeguard ensures that the effective cost due to limited visual acuity is slightly greater than one time step per character space deviation from the optimal viewing position. Also, because lexical processing does not continue during the saccade, the duration of the saccade (one time step) is not subtracted from the remaining time that is needed to identify the word. Thus, in this example, five more time steps of processing would still be needed to identify the word from the second viewing position (instead of four).

Finally, it is important to acknowledge that our assumption about visual acuity was not expected to permit the agent to simulate the inverted optimal-viewing position (IOVP) effect, or the finding that during natural reading, the initial fixation on a word tends to be of shorter duration if it lands on either the beginning or the end of the word than in the middle of the word (Vitu, McConkie, Kerr, & O'Regan, 2001). To account for this phenomenon, the agent may require additional assumptions, such as the ability to make rapid corrective saccades to move the eyes to a better viewing locations in cases involving misplaced fixations (Nuthmann et al., 2005). This assumption has been incorporated into the most recent version of the SWIFT model of eye-movement control in reading to provide an account of the IOVP effect (Engbert et al.,

[4] It was important to keep the simulations as simple as possible because the value functions used in our reinforcement-learning simulations were implemented as look-up tables. Because this requires a large amount of RAM, simulating sentences longer than those presented in this article requires the simulations to be run by pooling RAM from several different computers, which makes the simulations costly (in terms of time) because of limits on intercomputer network communication speed. We are currently investigating alternative methods of implementing the algorithm (e.g., replacing the look-up tables with function approximation using residual algorithms that are implemented in neural networks; Baird, 1995) so that in the future, larger simulations can be run on conventional computers.

in press). However, given that our primary goal is to better understand the factors that contribute to the development of eye-movement behavior during reading on a fairly coarse level, and given that the IOVP effect is only one of many phenomena that must be explained by any complete theory of eye-movement control, we decided that the present assumption about visual acuity was adequate to use as a starting assumption.

Attention

Our final assumption was that attention is allocated serially during reading to only one word at a time. In our simulations, the serial processing of words was instantiated by allowing the agent to process only one word at a time; that is, the agent could not begin processing word $n + 1$ until after word n had been identified. The shifting of attention was assumed to take one time step, with no lexical processing occurring during the shifts.

This final constraint that only one word can be processed at a time is debatable. For example, the results of several experiments have been interpreted as evidence that lexical processing can be done on more than one word at a time (Inhoff, Eiter, & Radach, in press; Inhoff, Starr, & Shindler, 2000; Kennedy, 1998, 2000; Kennedy, Murray, & Boissiere, 2004; Kennedy, Pynte, & Ducrot, 2002; Murray, 1998; Vitu, Brysbaert, & Lancelin, 2004), and several existing models of eye-movement control also assume parallel lexical processing (Engbert et al., 2002; Reilly, 1993; Reilly & Radach, 2003; Yang & McConkie, 2001, 2004).

However, there are valid reasons to question studies that purportedly show parallel lexical processing (Rayner & Juhasz, 2004; Rayner, White, Kambe, Miller, & Liversedge, 2003). Thus, the alternative assumption is also supported by empirical data and is consistent with the fact that word identification is a difficult, attention-demanding process. For example, there is evidence suggesting that the individual visual features of each word must be attended to so that they can be bound together into a single representation, and that the attention that is required for this is limited and can be focused on only one word object at any given time (Treisman & Souther, 1986; Wheeler & Treisman, 2002; Wolfe, 1994; Wolfe & Bennett, 1996). Serial processing may also be necessary to avoid potential cross-talk between the lexical codes (orthographic, phonological, and semantic) of adjacent words. One final advantage of serial processing is that it presumably helps the reader maintain the order of the words in a sentence. This is useful in languages like English because word order conveys a considerable amount of syntactic information (Pollatsek & Rayner, 1999). And even in highly inflected languages, like Finnish, Turkish, and Hungarian, word order conveys important, relevant information about discourse status (Vallduvi & Engdahl, 1996) and topic focus (Kaiser & Trueswell, 2004). We therefore believe that these factors make it very unlikely that readers normally process more than one word at a time in natural reading.

Learning

Given these constraints, let's now consider how the agent learns to move its eyes. The adaptive reading agent was trained using the value-iteration reinforcement-learning algorithm (Coulom, 2002). Although there are many learning algorithms that permit adaptive systems to approximate functions to any desired degree of accuracy (e.g., back-propagation can be used to train neural networks; Rumelhart, Hinton, & Williams, 1986), many of these techniques can be used only if the desired output for each input is already known. Such algorithms fail to satisfy the demands of the present research problem because we do not, for example, want to make any a priori assumptions about what specific actions need to be executed or when these actions need to be executed for the agent to become an efficient reader. Reinforcement-learning algorithms avoid this problem by specifying how a low-dimensional error signal that results from each of the agent's actions can be integrated into the learning process so that the agent can develop an optimal strategy to maximize the total reward (or minimize the total punishment) that it receives.[5] From this perspective, the task that the agent faces is one of learning which action to perform (e.g., move the eyes) from whatever state the agent is in (with the current state being defined by the word being fixated, how much lexical processing has been completed, and so forth) to maximize the cumulative reward that it receives.

Table 1 shows the nine dimensions that define the set of all possible states that the agent can be in at any given time. Seven of these dimensions are represented by integer values (e.g., the length of the currently attended word is represented in character spaces); the other two are represented by Boolean (true/false) values (e.g., at any given time, a saccadic program either has or has not been initiated). During learning, the agent can use any of the state information that is available at that time to "decide" which action(s) to execute next. This state information is either sensory (coming from external sources; e.g., the length of the currently attended word) or cognitive (represented internally by the agent; e.g., whether a saccadic program has been initiated).

During each simulation time step (see Footnote 2), the agent can execute one or more possible actions: (a) continue processing the word that is currently being attended, and optionally (b) initiate a saccadic program to move the eyes a particular number of character spaces, or (c) shift attention to the next word. (There is a fourth action that can only be executed after the last word in a sentence has been identified; this action terminates the simulation trial.) These actions have consequences for the agent; it is rewarded for each word that is identified (i.e., +1 is added to the

[5] Reinforcement learning has been proposed as a model of at least some kinds of reward-based motor learning (Barto, 1995; Schultz, 1998; but see also Wörgötter, in press). A considerable amount is also known about the biology of reinforcement learning, which involves dopaminergic (DA) cells in the ventral tegmental area and the substantia nigra pars compacta. These DA cells signal discrepancies between expected and received reward coming from particular sensory inputs; the DA cells become active with unanticipated rewards and are inhibited when anticipated rewards are not received (Schultz, Dayan, & Montague, 1997). We propose that the striatum (which receives convergent input from cerebral cortex and DA cells) may represent the relative values of different sensory and cognitive states, and that the caudate nucleus may provide mappings between these states and possible motor actions by projecting to output nuclei in the basal ganglia, including both the global pallidus internal segment and the substantia nigra pars reticulata. The latter area directly influences the superior colliculus and may be responsible for initiating saccadic programming (Hikosaka, Takikawa, & Kawagoe, 2000). However, it is important to note that we are not making strong claims about the biological plausibility of the value-iteration algorithm that was used in our simulations; we simply view it as being a way to examine how sophisticated eye-movement behavior might emerge through the task constraints that readers face in learning how to read efficiently.

Table 1
Nine Dimensions That Define the Set of States for the Adaptive Reading Agent

Dimension #	State information	Information source (information type)
1	Has the current word been identified?	Cognitive (Boolean)
2	How many time steps have already been spent processing the attended word?	Cognitive (Integer)
3	How many character spaces are between the fixation location and the center of the word being processed?	Sensory (Integer)
4	What is the length of attended word (i.e., word *n*)?	Sensory (Integer)
5	What is the length of next word (i.e., word $n + 1$)?	Sensory (Integer)
6	What is the length of previous word (i.e., word $n - 1$)?	Sensory (Integer)
7	Is a saccade currently being programmed?	Cognitive (Boolean)
8	What is the length of saccade that is being programmed?	Cognitive (Integer)
9	How many time steps have already been spent programming a saccade?	Cognitive (Integer)

reward that it receives), and it is punished for each time step that it spends identifying a given word (-1 is added to the total reward). The goal of the agent is to learn the optimal mapping between sensory inputs and internal states, on one hand, and permissible actions, on the other, that will maximize the overall reward that it receives. We predicted that this method of shaping behavior, in conjunction with the constraints that were already described, would allow the agent to learn to move its eyes in a way that would maximize its overall reading rate.

The mapping that the agent learns between the state information and the set of possible actions is specified by an optimal-value function, V(x). This function is derived from the value-iteration algorithm (which is based on temporal-difference learning; Barto, 1995) and is specified as follows:

Value-iteration algorithm:

$i = 0$

for all S: $V_i(S) = 0$

repeat:

$i = i + 1$

for all S: $V_i(S) = V_{i-1}(S)$

$+ \varepsilon\{\max_{action \in M}[\text{reward}(S, \text{action})$

$+ \gamma V_{i-1}(S')] - V_{i-1}(S)\}$

until learning has completed.

During each iteration of learning (indexed by i), the agent evaluates the reward that will result from executing each of the set of M actions that are permissible from each state S that the agent can be in. The agent selects the action that maximizes the sum of the immediate reward (the left-hand term in the square brackets, above) and some fixed proportion ($\gamma = 1$) of the reward that the agent predicts that it will receive from the state (S') that results from this action based on the value that was associated with that state during the previous learning iteration ($V_{i-1}[S']$). A learning-rate parameter ($\varepsilon = .1$) can be used to attenuate the agent's learning rate. Before we discuss the reason for attenuating learning, however, we will first provide a simple example to show how the value-iteration algorithm works. This example is schematically depicted in Figure 1.

Figure 1 depicts a simple two-dimensional world that an agent must learn to traverse, moving from the start state to the goal state. From each location or state (depicted as circles labeled A-I), the agent can perform only a limited number of actions (depicted by arrows): move to an adjacent location or—from within the goal state—remain in the same location. The rewards that are associated with each action are indicated by the numbers beside the arrows; similarly, the values that are associated with each state are indicated by the numbers inside the circles. Notice that all of the actions except for the one of remaining in the goal state are punished (i.e., reward $= -1$). This means that the agent can maximize the amount of reward that it receives (i.e., it will minimize its punishment) by traversing the world as quickly as possible, using the minimum number of actions to move from the start state to the goal state. How does this agent learn to do this? The initial value function and three iterations of learning (starting in the top panel) are shown to illustrate how the value-iteration algorithm allows the agent to learn this task.

Prior to any learning (i.e., $i = 0$; top panel), all of the states have equal values, that is, for each state, $V_0(S) = 0$. During the first learning iteration ($i = 1$; second panel), the agent evaluates the outcome of each possible action that it can execute from each state and then updates the value associated with each state using these predicted outcomes. For example, the agent predicts that the best action to take from the goal state (S_C) is to simply remain in that state; this action is not punished (reward $= 0$), and the state that results from this action (S') is predicted to have a value of zero associated with it. (This prediction is based on the value of the goal state at $i = 0$; i.e., $V_0[S'] = V_0[S_C] = 0$.) Thus, by remaining in the goal state, the agent predicts that it will minimize the overall punishment that it will receive (which happens to equal 0), as follows:

$$V_1(S_C) = V_0(S_C) + .1\{\max_{action \in M}[\text{reward}(S_C, \text{action}) + 1 \times V_0(S')] - V_0(S_C)\}$$

$$V_1(S_C) = V_0(S_C) + .1\{[\text{reward}(S_C, \text{“remain in } S_C\text{”}) + 1 \times V_0(S_C)] - V_0(S_C)\}$$

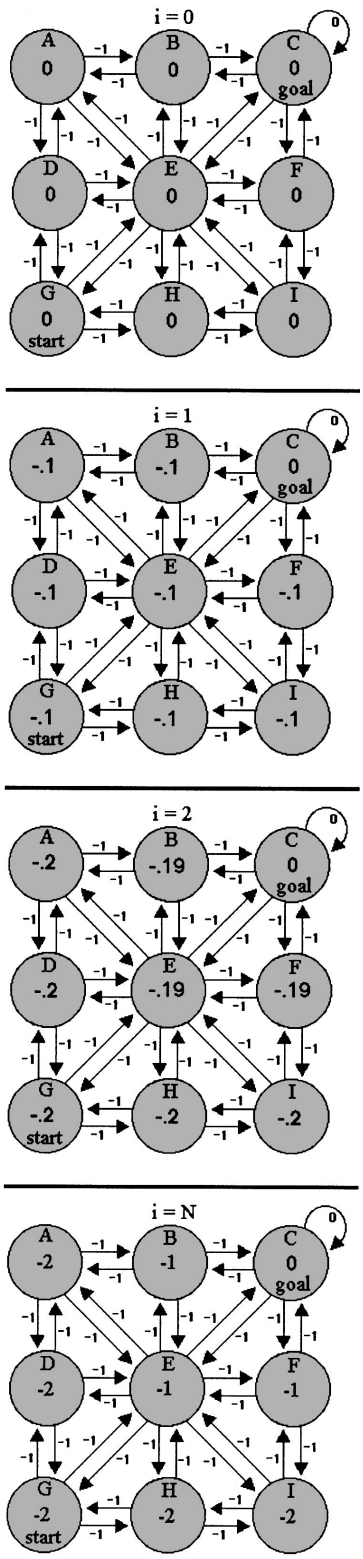

Figure 1. An example illustrating how the value-iteration algorithm works. In this example, the agent must learn to traverse a two-dimensional "world," moving from the start state to the goal state as quickly as possible. The top panel shows the initial value function, and the other three panels show the value function at three different stages of learning (e.g., the bottom panel shows the value function after *N* iterations of learning). The circles represent possible states or locations (labeled with letters A-I) that the agent can occupy, and the arrows represent the set of actions (in this example, movements) that the agent can make from each state. Each action has a reward associated with it (the numbers next to the arrows), and each state has a value associated with it that can change with learning (the numbers in the states). In this example, the agent is punished (i.e., it receives a reward of −1) for each action that it performs until it moves into the goal state. (The action of remaining in the goal state is not punished, as indicated by the special action that has a reward equal to 0.)

$$V_1(S_C) = 0 + .1\,[(0 + 1 \times 0) - 0]$$

$$V_1(S_C) = 0$$

The agent then updates the value of the goal state with the value that it predicts will be associated with the goal state if it remains in that state. This new value is shown in the second panel, that is, $V_1(S_C) = 0$. The agent then repeats this process for all of the other states, evaluating the set of possible actions from each state and updating the value associated with each state that is predicted to result from taking the action that produces the maximum reward. The second panel shows the values that result from the application of this algorithm after the first learning iteration (i.e., i = 1).

The process of evaluating the outcomes of the set of possible moves from each state continues during the second learning iteration (i = 2; third panel). For example, from the right middle location (S_F), the agent decides that the best action—although all three of the possible actions will be punished to the same degree—is to move to the goal state because this location is predicted to be associated with the least future punishment, that is, $V_1(S') = V_1(S_C) = 0$. Thus, from the right middle location, the agent predicts that the action of moving to the goal state will minimize the punishment that it will receive, as follows:

$$V_2(S_F) = V_1(S_F) + .1\,\{\max_{action \in M}[\text{reward}(S_F, \text{action}) + 1 \times V_1(S')] - V_1(S_F)\}$$

$$V_2(S_F) = V_1(S_F) + .1\,\{[\text{reward}(S_F, \text{“move to } S_C\text{”}) + 1 \times V_1(S_C)] - V_1(S_F)\}$$

$$V_2(S_F) = -.1 + .1\,[(-1 + 1 * 0) - (-.1)]$$

$$V_2(S_F) = -.19$$

After the second iteration of learning is complete, the agent has already learned enough information about its world for it to move from the start state—or any state—in a straight line toward the goal state. The information, or *policy*, that allows the agent to do this is based on the value function, $V_2(S)$, which contains the values that were associated with each state after the second iteration of learning (i = 2; third panel). To make use of the policy, the agent can move toward the goal state from any other state by repeatedly selecting and executing those actions that minimize the amount of punishment that it expects to receive from each new state. For example, from the start state (S_G), the agent will first move to the central location (S_E) because its value is less negative ($V_2[S_E] = -.19$) than those of the other two locations that the

agent can move to ($V_2[S_D] = V_2[S_H] = -.2$). From the central location (S_E), the agent will then move to the goal state (S_C) because this location is also associated with a less negative value ($V_2[S_C] = 0$) than any of the other permissible locations (which range from $-.19$ to $-.2$).

Finally, after some criterion of learning has been reached, the training process is stopped. The bottom panel shows the final policy that results after an additional N learning trials ($i = N$; bottom panel). In this deterministic world, where the results of each action are predictable, the values of the value function will eventually converge to yield the optimal-value function. Because the reward values in this particular problem were equal to -1 for each action or time step, each state's value will approach a value equal to the number of time steps that are necessary to move from that state to the goal state.[6] For example, the agent requires two time steps to move from the top left state (S_A) to the goal state (S_C).

The value-iteration algorithm can thus be described by the following simple dictum: It is better to take actions that result in states from which more positive reward can be obtained, and it is better to avoid taking actions that result in states from which relatively less reward can be obtained. In the simulations described below, the agent did not learn to move through a two-dimensional world but instead learned to move through the state space defined by the nine dimensions displayed in Table 1. To do this, the agent had to learn when and where to move its eyes so as to identify sequences of words as quickly as possible. As mentioned, the agent was punished -1 for every time step that it spent processing the sentence and was rewarded $+1$ for each word that was identified. These values are largely arbitrary in that they do not affect the policy that is learned but will instead only affect how long it takes to learn the policy. After the agent learns to perform this task, the policy for how it will read a given sentence can then be derived from the optimal-value function by placing the agent in the starting state, selecting the action that leads to the next most highly valued state, moving to that state, and then repeating this process (selecting the action that leads to the next most highly valued state) until a terminal state is reached. In these simulations, the starting state is the state where the first letter of the sentence is being fixated, no saccade is being programmed, and no time has been spent identifying the word that is being attended. The terminal states are simply those in which the last word in each sentence has been identified.

In the two simulations that are reported below, the free parameter γ was set equal to a value of 1 so that the immediate reward that the agent would receive from any action would be weighted as much as the future reward that it would receive from the resulting state. This was done to maximize the effects of the reward that the agent would receive from future states. (Values of γ less than 1 are sometimes used to avoid situations in which the agent might expect an infinite amount of reward because of actions that might move it back to earlier states, thereby allowing the agent to obtain even more reward than it would from moving to the terminal state. Our assumption that attention could only shift from left to right was sufficient to prevent this type of situation from occurring in our simulations.)

Finally, as already mentioned, the free parameter ε can be set equal to a value of less than 1 to attenuate the learning rate and thereby ensure that the values that are associated with the states during one learning iteration will not simply be replaced by the values that are associated with those states during the next iteration. It was necessary to do this in our second simulation because the introduction of stochastic saccadic error would otherwise cause the agent to erroneously associate the outcomes resulting from misplaced fixations with the actions that the agent had intended to execute (i.e., the agent would associate the reward that resulted from the saccade that was actually executed with the saccade that was intended). By using $\varepsilon = .1$ in our second simulation, the agent was able to learn the average outcome that resulted from several attempts to make saccades of a given length from each state. Because our first simulation was deterministic and did not include saccadic error, the value of ε was set equal to 1 in that simulation. Although we could have assessed the effects of saccadic error using a single learning rate parameter (i.e., by setting the value of ε equal to .1 in Simulation 1), we did not do so because ε in a deterministic problem affects only the agent's rate of learning, not its behavior after learning has reached asymptote. The value of ε was therefore set equal to 1 in Simulation 1 to ensure a maximal rate of learning.

Simulation Results

In the two simulations that are described below, the agent was trained on the same set of 20 sentences (see Table 2). Each of these eight-word sentences began and ended with one-letter words (which were both excluded from all of the analyses because processing begins and ends abruptly on the first and last words of each sentence) and contained a different random permutation of 1-, 3-, 5-, and 7-letter words. The agent was trained on these different sentences to ensure the generality of the results by minimizing the contributions of any idiosyncratic effects that may have resulted from using only a single sentence (e.g., skipping that may result from a sentence containing two adjacent short words). The reported simulation results thus reflect the agent's average behavior across the 20 different sentences. The results of these simulations are reported next.

Simulation 1

In the first simulation, we were mainly interested in whether the agent would learn to produce eye movements that—on some coarse level—resembled those that are observed with real human readers. The specific questions that we intended to examine included: (a) Would the agent learn to fixate each word exactly once, or would it instead occasionally skip or refixate some words? (b) Would the agent learn to skip the shorter, easier-to-identify words

[6] Value-iteration is not guaranteed to produce an optimal-value function in nondeterministic learning environments. The practical consequence of this is that in our second simulation (which introduced stochastic saccadic error), the optimal-value function did not converge. We therefore stopped training the agent after 100 iterations of learning because changes in the agent's behavior were negligible after this amount of training. One solution to ensure convergence might be to alter the value-iteration algorithm so that the value of any future state is the expected value of the future state given some amount of sampling of the stochastic environment that results from each possible action. Future work will be needed to explore the viability of this approach.

Table 2
Twenty 8-Word Sentences Used in Simulations 1 and 2

Sentence #	Sentence
1	1, 1, 5, 3, 7, 5, 3, 1
2	1, 1, 7, 3, 5, 7, 5, 1
3	1, 5, 3, 1, 5, 1, 1, 1
4	1, 3, 5, 3, 5, 1, 3, 1
5	1, 1, 3, 5, 7, 7, 5, 1
6	1, 7, 1, 7, 3, 7, 5, 1
7	1, 1, 3, 1, 7, 5, 3, 1
8	1, 7, 3, 7, 7, 5, 5, 1
9	1, 3, 3, 1, 7, 7, 3, 1
10	1, 3, 1, 3, 5, 7, 3, 1
11	1, 7, 1, 7, 7, 1, 3, 1
12	1, 7, 3, 1, 1, 5, 3, 1
13	1, 5, 3, 5, 3, 7, 1, 1
14	1, 3, 7, 5, 7, 5, 3, 1
15	1, 7, 7, 1, 3, 1, 5, 1
16	1, 5, 1, 5, 5, 7, 3, 1
17	1, 1, 3, 7, 3, 1, 3, 1
18	1, 5, 7, 5, 5, 7, 1, 1
19	1, 3, 1, 5, 3, 3, 5, 1
20	1, 5, 3, 5, 5, 7, 5, 1

Note. The numbers in the right-hand column indicate the lengths of the words in the sentences; for example, Sentence 1 consisted of a one-letter word, followed by another one-letter word, followed by a five-letter word, a three-letter word, and so on.

more often than the longer, more difficult words, and would it learn to refixate difficult words more often than easy words? (c) Would the agent learn to spend more time fixating the longer, more difficult-to-identify words than the shorter, easier-to-identify words? (d) Would the agent learn to direct its saccades toward the optimal viewing positions of words? (e) Would the agent learn to initiate saccadic programs to move the eyes from word n to word $n + 1$ before it had completely identified word n? To address these questions, the agent was first trained for 100 learning iterations (i.e., the number of iterations that were needed before the optimal-value function converged) on each of the 20 sentences in our corpus (see Table 2). The agent was then evaluated on those same sentences by first reinstating the start state for each sentence and then allowing the agent to execute whatever sequence of actions it had learned to perform so as to read the sentence as quickly as possible. This training and testing procedure was done separately for each sentence because of the restrictions discussed in Footnote 4. Although this might raise concerns about how well the simulation results might be expected to generalize across different (novel) sentences, all of the qualitative patterns that are presented below were extremely robust and were consistently observed across all of the sentences used in both the current simulations and in several preliminary simulations that are not reported in this article.[7]

The results of Simulation 1 indicated that the agent learned to fixate the center of each of the words in the sentence corpus exactly once, and that the agent did not skip words, refixate words, or make regressions. (This pattern was evident for words of all lengths, i.e., all levels of word-processing difficulty.) This result is perhaps not surprising given that the agent was indirectly penalized by increasing the word-identification time by one time step for every character space of distance between the fixation location and the center of the word being processed. (This was done to simulate the slowing in lexical processing that results from limited visual acuity; Lee et al., 2003; Rayner & Bertera, 1979; Rayner & Morrison, 1981.) This cost was clearly sufficient to affect the agent's eye-movement behavior, causing the agent to learn to always direct its eyes toward the optimal viewing position. The agent was always able to do this because of our simplifying assumption that saccades are executed without error. (The effects of motor error on saccadic accuracy are examined in the next simulation, reported below.) Thus, in the absence of saccadic error, the optimal solution for the agent is to fixate the center of each word exactly once.

The agent was also sensitive to word-processing difficulty and spent less time looking at the shorter, easy-to-identify words than the longer, more difficult words. (Remember that in our simulations, 1-, 3-, 5-, and 7-letter words required 4, 5, 6, and 7 times steps to identify, respectively.) The agent learned to move its eyes so that each word was fixated for exactly the amount of time that was needed to identify the word (e.g., the agent spent four time steps looking at one-letter words, which took exactly four time steps to identify). This result indicated that the agent was able to learn the positive correlation between a word's length and its processing difficulty. Moreover, the agent was able to anticipate how long it would take to identify word n and then use this information to start programming a saccade to word $n + 1$ so that the program (which always took three time steps to complete) finished just as word n had been identified. The policy that was learned by the agent can thus be described as one of anticipating how long it took to identify each word and then initiating a saccadic program so as to keep its eyes on the word just long enough for it to be identified.

In summary, the agent in Simulation 1 learned a policy of always fixating each word exactly once, from each word's optimal viewing position, and initiating saccadic programs to move its eyes from one word to the next so that each word was fixated just long enough to be identified. This policy was evidently the optimal solution for identifying the words in each of the sentences as rapidly as possible and in their correct order, given the constraints that had been placed on the agent (e.g., limited visual acuity).

Simulation 2

Our second simulation was intended to replicate the main findings from the first but with the added assumption that saccades are subject to motor error. To add saccadic error, a random deviate was sampled from a uniform distribution and then normal approximation was used (truncating to avoid values more extreme than $\pm$ 3 standard deviations from the mean and rounding to the closest character space) to determine the amount of saccadic error that would be added to the length of the saccade that the agent intended to execute. The standard deviation of the saccadic error was equal

[7] One of these simulations used a corpus of 25 three-, five-, and seven-letter words and saccadic error that was sampled from a uniform distribution (Reichle & Laurent, 2004). Two other simulations were identical to the simulations that are reported in this paper except that they used slightly shorter saccadic programming times (two time steps), a smaller corpus of 10 six-word sentences, and a different number of learning iterations (60 without saccadic error and 750 with saccadic error).

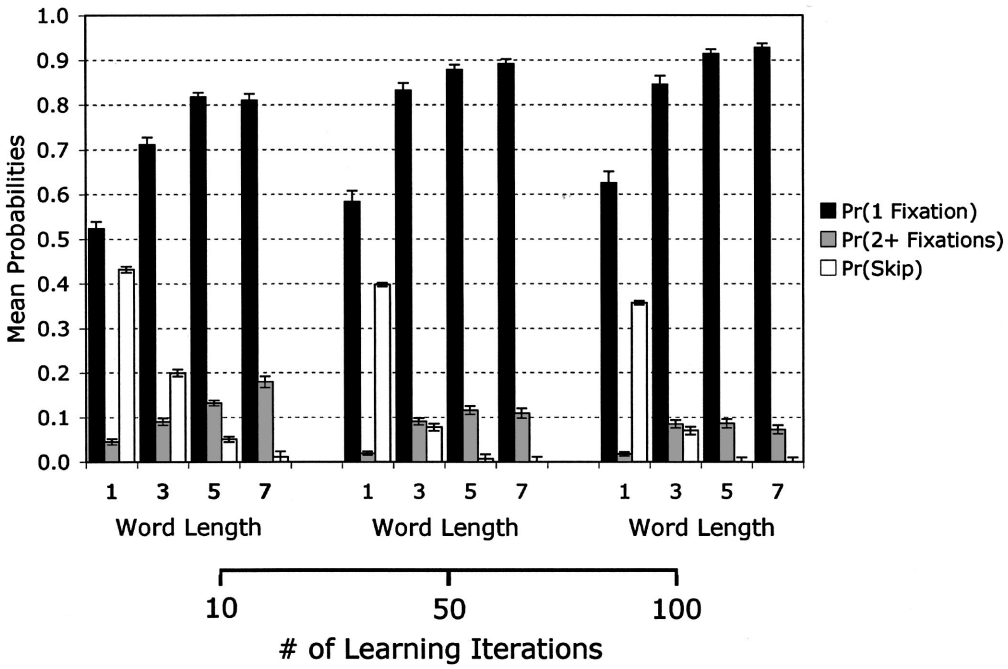

Figure 2. Simulation 2: Mean probabilities of making single fixations, skipping, and making multiple fixations on 1-, 3-, 5-, and 7-letter words, after three levels of learning (10, 50, and 100 learning iterations). The bars show the standard errors of the means.

to one character space, so that the amount of saccadic error extended over the range of −3 to +3 character spaces.[8]

The new agent, now subject to saccadic error, was trained for 100 learning iterations on each of the 20 sentences that were used in Simulation 1. The same sentences were used to allow direct comparisons between the two simulations. The agent was evaluated as in Simulation 1: The starting state for each sentence was first reinstated, and then the agent was allowed to select whatever sequence of actions it had learned to execute to read each sentence as rapidly as possible. To examine the effects of learning, we saved and evaluated the agent's policies after 10, 50, and 100 learning iterations.

Although our decision to examine the effects of learning after 10, 50, and 100 learning iterations (instead of after some other number of learning iterations) was somewhat arbitrary, it allowed us to describe the agent's behavior and how it changed across the full range of learning. Another basis for our decision was that it was not possible to examine the agent's behavior after fewer than 10 learning iterations because with fewer than 10 iterations, the agent sometimes behaved erratically (e.g., it would not move its eyes but would instead identify all of the words in a sentence by only shifting its attention). This erratic behavior was due to the fact that with fewer than 10 iterations of learning, the agent did not have enough experience to know that this type of behavior was not optimal. In addition, this erratic behavior is clearly very different from the eye-movement behavior that is typically observed with children who are learning how to read (Rayner, 1978). This discrepancy highlights one important limitation of using an ideal observer to understand human performance: The simulated trajectory of learning does not perfectly mirror the trajectory that is observed with humans.[9]

Finally, because saccadic error also occurred during our evaluations of the agent's behavior, the behavior was evaluated over 50 Monte Carlo encounters with each sentence. In the analyses reported below, the mean performance across the 20 sentences was first calculated for each of the 50 Monte Carlo runs of the agent. Inferential statistics were then calculated using these means. (We opted to perform these subject analyses using the Monte Carlo runs as the units of analysis to avoid the problems associated with missing values; e.g., a few of the sentences did not contain words of a given length. However, item analyses using the sentences as the units of analysis produced nearly identical patterns of results.)

Figure 2 shows the mean proportion of times that the agent

[8] The results of several eye-tracking experiments indicate that there are two independent sources of saccadic error during reading: a systematic error, which is a function of how much the intended saccade target deviates from some optimal saccade length (which is approximately seven character spaces in English), and a random error component, which tends to increase with saccade length (McConkie et al., 1988, 1991; Rayner, 1977; Rayner et al., 1996). Although including these two sources of saccadic error would have made our simulation more realistic, it would have also made it much more complex.

[9] It is important not to draw too strong of a parallel between the agent's behavior and children who are learning to read. The reason for this disclaimer is that children who are learning to read must learn to perform a variety of reading-related skills, including identifying words. In contrast, the agent can already identify words but must learn to anticipate how long it will take to identify each word (using word length) and then learn when and where to move its eyes so as to make the process of identifying words as efficient as possible.

made single fixations, skipped, and made multiple fixations on 1-, 3-, 5-, and 7-letter words after 10, 50, and 100 iterations of learning. In contrast to what was observed in the first simulation, the figure shows that the agent skipped some words and refixated others. We evaluated these dependent measures with analysis of variance (ANOVA), using word length and the number of learning iterations as within-subjects factors. These analyses indicated that longer words tended to be the recipients of multiple fixations and to be skipped less often than shorter words (both Fs > 44; both ps $< .001$). These analyses also indicated that the effect of learning was reliable with both dependent measures (both Fs > 30; both ps $< .001$), and that learning interacted with word length on both measures (both Fs > 6; both ps $< .001$). As Figure 2 indicates, the agent learned to fixate most words exactly once but also learned to skip short words and to refixate long words.

Although many of the skips and refixations that are evident in Figure 2 were caused by saccades that either overshot or undershot their intended word targets because of simple motor error, a significant proportion of both the skips and refixations were deliberate in the sense that they were part of the policy that was learned by the agent. For example, the agent often skipped a short word if it was to the right of another short word (e.g., in Sentence 12, the agent sometimes skipped the one-letter word immediately to the right of the three-letter word; see Table 2). To quantify the agent's behavior, we calculated the percentage of the 6,000 words in Simulation 2 (i.e., 6 words per sentence $\times$ 20 sentences $\times$ 50 Monte Carlo runs) that the agent attempted to skip. This analysis indicated that after 10, 50, and 100 learning iterations, the agent attempted to skip 8.82%, 2.65%, and 1.3% of the words, respectively. However, because of saccadic error (which eliminated some of deliberate skips), the agent was only successful in deliberately skipping 4.87%, 1.08%, and 0.12% of the words. A similar analysis of the refixations indicated that although the agent attempted to refixate 9.25%, 7.95%, and 6.87% of the words at the three stages of learning, it was only able to successfully do so for 7.73%, 6.95%, and 5.83% of the words.

Although the agent produced too few deliberate skips and refixations to be an accurate description of real human performance (see Rayner, 1998), the skips and refixations were sufficient to show that these behaviors were adaptive in that they allowed the agent to perform its task in an optimal manner. In other words, the fact that the agent deliberately skipped some words and refixated others demonstrates that such behaviors are part of the policy that it learned to read as efficiently as possible. However, it is also important to point out that the agent's tendency to make fewer skips as it learned to move its eyes is seemingly at odds with the fact that children who are learning how to read typically come to skip more words as they become better readers (Rayner, 1978). One possible explanation for this apparent discrepancy is that in contrast to human readers, the agent does not have additional sources of information (e.g., a word's predictability within a given sentence context; Balota et al., 1985; Ehrlich & Rayner, 1981; Kliegl et al., 2004; Rayner et al., 2004; Rayner & Well, 1996) to help it identify words. If the ability to use higher order linguistic information improves with reading skill, then one might expect that at least some of the increased skipping that is observed with children who are learning to read results from their increasing ability to more effectively use this information to help them identify words. For example, by learning how to use predictability to identify words in the parafovea, skilled readers might be able to skip more words. Future simulations will be necessary to test this hypothesis.

Another interesting finding related to skipping was that although the deliberate skips allowed the agent to increase its overall reading rate, this skipping also resulted in local costs, inflating the fixation durations on the words immediately preceding and following the skipped words. To quantify this, we first identified all pairs of adjacent words from the 6,000-word corpus in which a given word (word n) was skipped at least once and the preceding word (word $n - 1$) was fixated only once. We also identified all pairs in which word n was skipped at least once and the following word (word $n + 1$) was fixated only once. We then calculated the mean fixation durations on words $n - 1$ and $n + 1$ as a function of whether word n was fixated, deliberately skipped because doing so was in the agent's policy, or accidentally skipped because of saccadic error. Although this method of examining skipping costs is less conservative than Kliegl and Engbert's (2005) method of examining skipping costs in consecutive word triplets (in which the words immediately preceding *and* following skips each received only a single fixation), our procedure was necessary to maximize the number of observations in our analyses.

Table 3 shows the number of fixations in each of the six conditions of interest, along with the means and standard deviations of the fixation durations in each condition. Table 3 also

Table 3
Number (N) of Fixations and Ms and SDs of Fixation Durations on Words n − 1 and n + 1 as a Function of the Number of Learning Iterations (10, 50, or 100) and Whether Word n Was Fixated, Accidentally Skipped, or Deliberately Skipped

Word	# of learning iterations	Word n fixated			Word n accidentally skipped			Word n deliberately skipped		
		N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
$n - 1$	10	3,676	6.85	4.03	459	7.29*	2.74	285	7.10	2.13
	50	3,264	6.59	3.89	411	7.51*	1.69	65	7.80*	2.51
	100	2,139	5.70	2.54	414	6.54*	1.49	7	8.00*	3.42
$n + 1$	10	2,620	6.63	3.89	437	8.00*	3.52	183	10.78*	4.79
	50	1,693	6.73	2.72	265	7.65*	2.58	42	9.14*	3.88
	100	1,149	5.96	2.27	247	6.85*	2.33	4	8.00	3.46

Note. * = the difference between the mean fixation duration on word $n + 1/n - 1$ when word n was skipped and the corresponding mean fixation duration on the same word when word n was fixated was statistically reliable ($p < .05$), using an independent-samples, two-tailed t test.

shows the results of several planned contrasts. The contrasts compared the mean fixation durations on words $n - 1$ and $n + 1$ in the conditions where word n was fixated with the mean fixation durations on words $n - 1$ and $n + 1$ in the conditions where word n was either deliberately or accidentally skipped.

Table 3 indicates that after 100 learning iterations, the mean fixation durations on word $n - 1$ were longer if the agent deliberately skipped word n than if it fixated word n (mean difference = 2.3 time steps). Table 3 also shows that this fixation-duration cost on word $n - 1$ was much smaller if the agent accidentally skipped word n (mean difference = .84 time steps). One possible explanation for the difference in the relative sizes of the cost before deliberate versus accidental skips is that the cost is largely incurred if the agent began processing the skipped word on the fixation prior to the skip, so that the time spent processing word n is added to the time spent fixating on word $n - 1$. This type of parafoveal processing would be more likely to occur before skips that are planned and less likely to occur before skips that resulted from saccades overshooting their intended targets. This explanation for why the fixation-duration costs on the launch-site word are not associated with all skips (but only with those that are deliberate) suggests why overall fixation-duration costs are sometimes found (Pollatsek, Rayner, & Balota, 1986; Pynte, Kennedy, & Ducrot, 2004; Rayner et al., 2004; Reichle et al., 1998) and sometimes not (Kliegl & Engbert, 2005; McConkie, Kerr, & Dyre, 1994; Radach & Heller, 2000): The relative proportion of deliberate versus accidental skips may modulate the overall amount of skipping cost that is observed in a given experiment. Finally, this explanation is also consistent with the fact that the absolute size of the fixation-duration cost on word $n - 1$ that resulted from deliberately skipping word n increased across the three levels of learning (mean differences = 0.25, 1.21, and 2.30 time steps after 10, 50, and 100 iterations, respectively), whereas the cost attributable to accidental skips remained relatively small across learning (mean differences = .44, .92, and .84). These trends suggest that the agent learned to selectively skip certain (mostly short) words, and that although this may have incurred local costs by inflating the fixation durations on the words immediately before the skips, these local costs were offset by the speed-up in the amount of time that was necessary to process the entire sentence.

Table 3 also indicates that after 100 iterations of learning, the mean fixation durations on word $n + 1$ also tended to be longer if the agent deliberately skipped word n than if it fixated word n (mean difference = 2.04 time steps). (This result must be interpreted with caution because word n was only deliberately skipped four times in this condition.) Similarly, there was a small but statistically reliable cost on word $n + 1$ if word n was accidentally skipped (mean difference = .89 time steps). The fact that both deliberate and accidental skipping of word n tended to inflate the fixation durations on word $n + 1$ is due to the fact that in both cases, any processing that was done on word $n + 1$ prior to the skip had to be completed from a more distant viewing location (i.e., word $n - 1$). Under such distant viewing conditions, word $n + 1$ is processed more slowly because of the assumption of limited visual acuity. Also, with accidental skips, the agent often had to continue processing word n from word $n + 1$, inflating the fixation durations on word $n + 1$. This account is supported by the fact that the fixation duration costs following skipping were statistically reliable for both accidental and deliberate skips at all three stages of learning (mean differences = .89 to 4.15 time steps). The fact that fixation-duration costs on word $n + 1$ were observed with both accidental and deliberate skips may also explain why overall fixation-duration costs following skips have been consistently reported in the literature (Kliegl & Engbert, 2005; Pollatsek et al., 1986; Rayner et al., 2004; Reichle et al., 1998).

Returning now to the issue of refixations, it is worth noting that the agent often deliberately refixated long words if the initial fixation happened to land on the beginning of the word. For example, in Sentence 17 (see Table 2), the agent sometimes deliberately moved its eyes to the middle of the seven-letter word if the initial fixation on the word happened to land on the beginning of the word. Such refixations allowed the agent to recover from misplaced fixations on the longer, more difficult-to-identify words by allowing these words to be processed from a second—and in most cases, better—viewing location (Vergilino & Beauvillain, 2000; Vitu et al., 2001).

Finally, the agent made a small number of interword regressions in Simulation 2. After 10, 50, and 100 iterations of learning, the agent moved its eyes back to 2.11%, 0.32%, and 0.12% of the words (respectively) in the 6,000-word corpus. Across the three stages of learning, these regressions occurred most often (97.39% of the time) after the agent accidentally skipped a word and then moved its eyes backward so that it could process the word from a better viewing position. Regressions are thus similar to refixations in that both types of behavior allow the agent to reduce the local (immediate) costs associated with processing words from poor viewing locations. However, the fact that only 2.61% of the skips that resulted in regressions were deliberate skips indicates that—in contrast to refixations—deliberate regressions were not often part of the agent's optimal behavior. Why is this the case? Interword regressions may be especially costly to the agent because any words that have to be identified to the right of the fixation following a regression will often have to processed from a distant viewing location, which slows parafoveal processing because of poor visual acuity.

The results of Simulation 2 thus indicate that the introduction of saccadic error did not eliminate the effect of word-processing difficulty that was observed in Simulation 1 but instead caused effects of word-processing difficulty to appear in other measures—the probabilities of skipping, making a single fixation, and making more than one fixation. We now turn our discussion to the fixation-duration measures.

Figure 3 shows the mean first-fixation durations (i.e., the duration of the first forward fixation on a word), gaze durations (i.e., the sum of all forward fixations on a word), and total viewing times (i.e., the sum of all fixations on a word, including those following regressions) on the 1-, 3-, 5-, and 7-letter words, again as a function of the number of learning iterations. Figure 3 also shows the mean times that the agent spent attending to (i.e., identifying) the words of each length as well as the amount of time that the agent spent attending each word before requesting a saccade to move its eyes to another word. As in Simulation 1, the agent spent less time fixating and processing the shorter, easier-to-identify words than the longer, more difficult-to-identify words. This pattern was evident for all three fixation-duration measures, and both measures of the time words were attended: ANOVAs indicated reliable learning and word-length effects, as well as reliable interactions between these two factors, for all five mea-

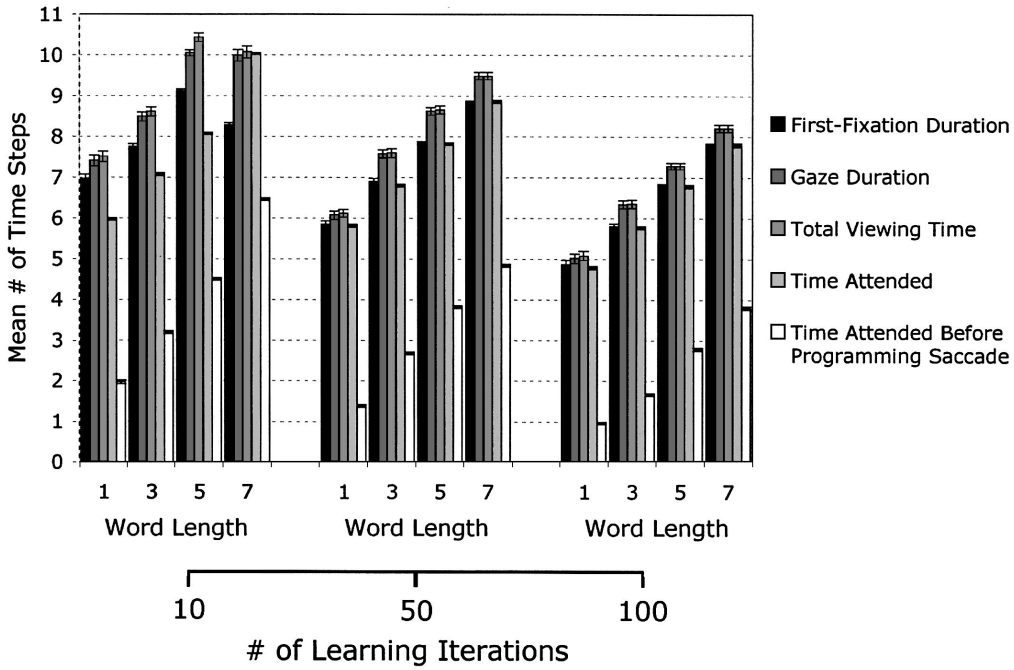

Figure 3. Simulation 2: Mean first-fixation durations, gaze durations, total viewing times, the mean times that the words were attended, and the mean times that the words were attended before initiating a saccadic program to another word (in arbitrary time steps) for 1-, 3-, 5-, and 7-letter words, after three levels of learning (10, 50, and 100 learning iterations). The bars show the standard errors of the means.

sures (all Fs > 15; all ps $< .001$). These results indicate that the agent learned the direct relationship between word length and word-identification difficulty and then used this information to anticipate how long it would take to identify the word. The agent then used this knowledge to begin programming a saccade to a new viewing location (to a character space in word $n + 1$ or word $n + 2$) prior to the time when the word n had been identified. One index of this anticipatory behavior is also shown in Figure 3: The white bars show the amount of time that the agent spent attending to word n before it started a saccadic program to move the eyes to word $n + 1$ or $n + 2$. After 100 iterations of learning, the agent initiated these saccadic programs an average of 3.99 time steps before word n had been identified (i.e., 3.99 time steps is the mean difference between how long the words were attended and how long the words were attended before the agent started programming a saccade). This result indicates that despite the uncertainty that resulted from the introduction of saccadic error, the agent was able to learn to use information about word-processing difficulty (i.e., information that covaried with word length) to decide when to move its eyes from one word to the next. As Figure 3 shows, the agent learned this behavior very rapidly; after only 10 learning iterations, the agent had already learned to start programming saccades to move the eyes from word n an average of 3.76 time steps before word n had been identified.

The fact that the agent rapidly learned when to initiate saccadic programs is somewhat paradoxical because the agent continued to learn the other important aspect of its behavior—where to move its eyes. This claim is supported by the fact that the average times that the agent spent fixating and processing words continued to decline across the three stages of learning. By learning where to move its eyes, the agent was able to spend a larger proportion of its time processing words from optimal or near-optimal viewing positions, which reduced the amount of time that the agent spent processing words from nonoptimal (distant) locations. The net effect of this was a reduction in both fixation times and word identification times (see Footnote 9).

Additional support for the claim that the agent continued to learn where to move its eyes can be seen in Figure 4, which shows the landing-site distributions for first fixations that were generated by the agent on 1-, 3-, 5-, and 7-letter words after 10, 50, and 100 iterations of learning. In contrast to what was observed in the first simulation, the landing sites show a considerable amount of variability, being normally distributed and centered over the middle letter (i.e., optimal viewing position) of the fixated words. This result is not unexpected (see Footnote 8), and it demonstrates that the general characteristics of the agent's eye-movement behavior (e.g., the relationship between word-processing difficulty and the fixation-duration measures) are fairly robust in the face of saccadic error. However, it is noteworthy that the landing-site distributions, although centered on the middles of the words, included a significant number of fixations on or near the beginnings of the words. For example, after 100 learning iterations, the agent was actually more likely to fixate the blank spaces immediately to the left of the seven-letter words than it was to fixate the first letters of the words: mean difference $= .048$; $t(49) = 6.49$, $p < .001$. This result was unexpected. Because the agent was penalized -1 for every character space between the fixation location and the center of the word being processed, this tendency to fixate near the beginnings

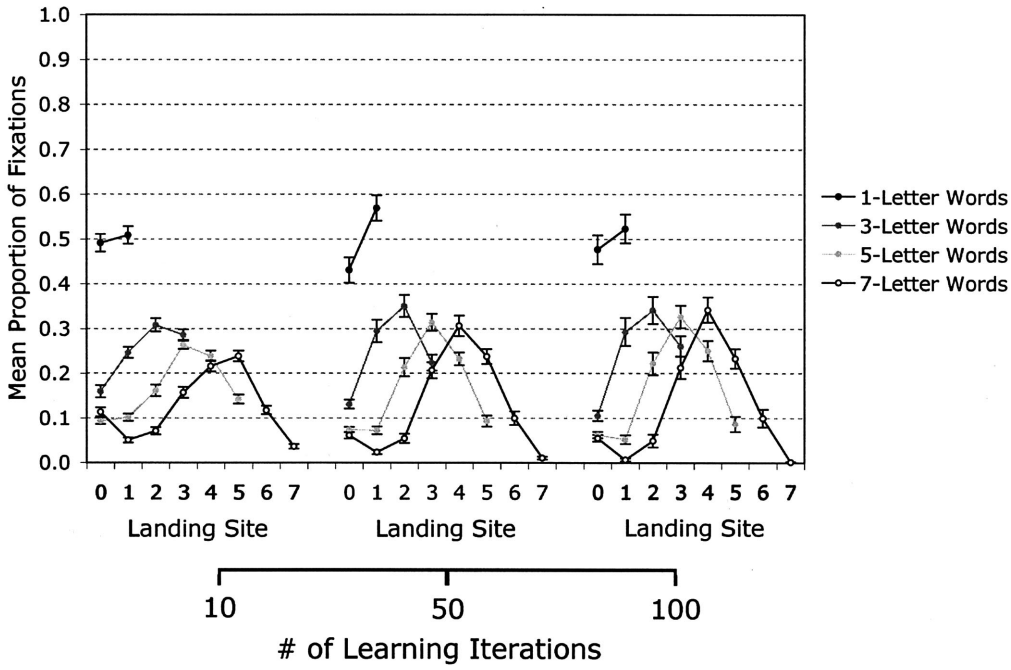

Figure 4. Simulation 2: Mean first-fixation landing-site distributions for 1–3-, 5-, and 7-letter words, after three levels of learning (10, 50, and 100 learning iterations). The bars show the standard errors of the means. The abscissa shows the character position of the landing site on a particular word, with 0 indicating the blank space immediately to the left of the word.

of longer words may reflect a balance that is struck between the potential speed-up in the overall reading rate that comes from making long progressive saccades and the potential cost associated with having to identify words from nonoptimal viewing positions because of motor error. This conservative strategy may help the agent to minimize its chances of having to make interword regressions, which as already discussed may be especially costly to the extent that they prevent efficient parafoveal processing. The agent's preference to incur a small local cost by occasionally fixating the beginning of long words (rather than their centers) may allow it to avoid the potentially larger costs associated with saccadic error or interword regressions. This possibility suggests that a similar strategy in humans may—at least in part—contribute to the finding that the peaks of the landing-site distributions of human readers tend to occur between the beginnings and the middles of a words (i.e., the preferred viewing location; McConkie et al., 1988, 1991; Rayner, 1979; Rayner et al., 1996). We will say more about this possibility in the General Discussion section.

General Discussion

The simulations described in this article show how complex patterns of eye movements can emerge from relatively simple task constraints that are imposed upon an agent who is trying to read efficiently. That is, the simulations show how fairly sophisticated eye-movement behavior can emerge from an agent that is trying to learn when and where to move its eyes so as to identify sequences of words as quickly as possible. In both simulations, the adaptive reading agent learned to direct its saccades toward the centers of words and to spend more time looking at the longer, more difficult-to-identify words. In the second simulation, the addition of saccadic error resulted in the agent learning to skip the shorter, easier-to-identify words and to refixate the longer, more difficult words. These behaviors are intelligent in that they allowed the agent to identify all of the words in the sentences as quickly as possible, and they are emergent in that the agent was not designed to exhibit them. And perhaps most important, these behaviors resemble those of skilled human readers (see Rayner, 1998), which suggests that we have implemented the important factors that influence the development of eye-movement behavior during reading—the task of identifying the words on the printed page rapidly and in order and the various physiological and psychological constraints that make this task difficult.

Our simulations also showed that the reading agent exhibited several other emergent eye-movement behaviors that may shed light on some unresolved issues related to eye-movement control during reading. The first of these behaviors is that the agent learned to anticipate how long it would take to identify words of each given length and then learned to start programming saccades to move its eyes from the words before they were completely identified (see Figure 3). This behavior was part of the agent's solution to the problem of learning when to move its eyes so as to identify all of the words in a given sentence in their correct order as rapidly as possible and under such constraints as limited visual acuity, saccadic error, and so forth. This solution is likely to be optimal because if the agent had learned to move its eyes from the words any later, this additional time would have accrued at each fixation location, thereby increasing the fixation durations and decreasing the agent's overall reading rate. Conversely, if the

agent had learned to move its eyes any earlier, then it would have moved its eyes from many of the words too soon and would have had to continue processing the words from a more distant viewing location (e.g., from a word to the right of the attended word). This would have slowed lexical processing because of the poorer visual acuity that is afforded by the more distant viewing locations, thereby resulting in longer fixation durations and decreasing the agent's overall reading rate. (This latter scenario may have made it necessary for the agent to make more regressions, which would also decrease its reading rate.)

What makes the agent's solution for deciding when to initiate saccadic programs especially interesting is that it is not necessarily the one that would be predicted. For example, several computational models of eye-movement control during reading posit an autonomous timer that is responsible for triggering the initiation of saccadic programs at random intervals, so that the eyes are moved through the text at a rate that is largely unaffected by lexical processing (Engbert et al., 2002, in press; O'Regan, 1990, 1992; Reilly & O'Regan, 1998; Reilly & Radach, 2003; Yang & McConkie, 2001, 2004). (These models do posit that difficulty in lexical processing can intervene and delay saccadic programming, thereby further modulating the random fixation durations and increasing the fixation durations on difficult-to-process words.) The agent could have adopted a similar strategy and simply learned to move its eyes at some random or fixed rate, occasionally pausing a bit longer to process especially difficult words (e.g., long words being processed from poor viewing locations). However, the agent did not adopt this kind of behavior. We believe that this type of behavior did not emerge in the agent because of the visual acuity constraint and because of the constraint that words had to be identified one at a time. The visual acuity constraint meant that the agent had to move its eyes to most of the words so that the words could be identified from the fovea (i.e., identified without the penalty that comes from poor visual acuity). The serial-processing constraint meant that the agent had to completely identify a word before it could begin processing the next. Given these constraints, the most efficient way for the agent to perform its task was to make its behavior contingent upon the moment-to-moment demands that it faced and to make its decisions about when to move its eyes from one word to the next contingent upon the processing difficulty of the words being fixated.

The more general question of when is the best time to start programming a saccade can be addressed by considering Figure 5. If one starts with the assumption that some aspect of the lexical processing of word *n* provides a signal to the oculomotor system to start programming a saccade to word $n + 1$ (Just & Carpenter, 1980, 1987; Pollatsek et al., in press; Rayner et al., 2004; Reichle et al., 1999, 2003, 2006; Reilly, 1993; Salvucci, 2000; Thibadeau et al., 1982), then it becomes important to understand which aspect of this processing actually triggers the programming. Figure 5 shows a time line of lexical processing and several possible links between lexical processing and saccadic programming. First note that some minimal amount of time is necessary for the visual information on the retina to be propagated to the brain. Recent evidence from several physiological experiments indicates that this eye-to-brain lag is approximately 50 ms in duration (Clarke, Fan, & Hillyard, 1995; Foxe & Simpson, 2002; Mouchetant-Rostaing, Giard, Bentin, Aguera, & Pernier, 2000; Van Rullen & Thorpe, 2001). This lag provides an absolute lower limit on when lexical

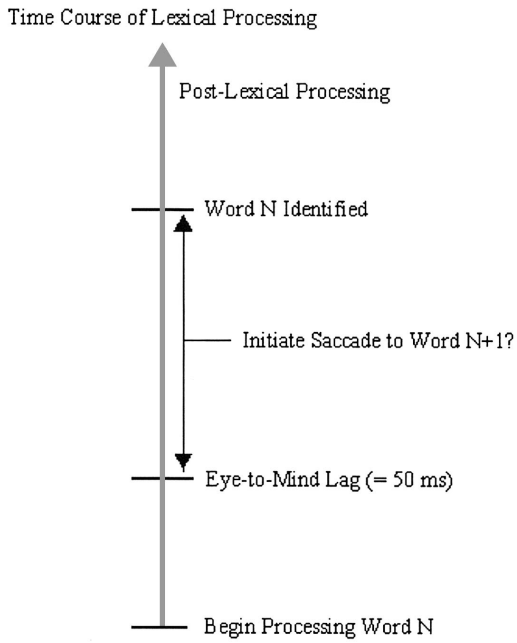

Figure 5. A schematic diagram showing the time course of lexical processing and saccadic programming.

processing might trigger saccadic programming; in other words, it takes at least a minimum of 50 ms for the visual information about a word to reach the brain and trigger saccadic programming. Given this restriction, what aspects of lexical processing might trigger saccadic programming?

One hypothesis is that the oculomotor system starts programming a saccade to move the eyes from word *n* to word $n + 1$ as soon as the lexical processing of word *n* begins. Unfortunately, this heuristic for deciding when to move the eyes is probably not going to be a good one because initiating a saccadic program to move the eyes off of word *n* as soon its processing begins will probably result in the eyes leaving the word too soon. To the extent that this happens, word *n* will have to be processed from a more distant viewing location, where its identification would be both slower and more prone to error (and probably result in too many regressions.)

Another hypothesis is that the program to move the eyes off of word *n* is initiated as soon as the word's meaning become available (i.e., lexical access). This solution avoids the pitfalls associated with moving the eyes too soon, but it is problematic for the opposite reason: The eyes have to remain on word *n* for the sum of the total time that is needed to identify the word (150–300 ms) and then program a saccade (125–150 ms), resulting in fixation durations that are considerably longer (275–450 ms) than those that are typically observed (200–300 ms; Rayner & Pollatsek, 1989).

One final hypothesis is that an intermediate stage of lexical processing (indicated in Figure 5 by the double-headed, black vertical arrow) is the trigger to start programming a saccade. This solution avoids the problems associated with moving the eyes too soon or too late and happens to be the one that was adopted by our reading agent: The agent learned to predict, from information available early in word identification, how long it would take to

identify word *n* (using its length, which was positively correlated to word-processing difficulty) and then used this prediction to decide when to begin programming an eye movement to word $n + 1$. Although our agent was allowed to use only the correlation between word length and processing difficulty to make its predictions, one might imagine that human readers use a variety of cues (e.g., activation of the word's orthographic or phonological codes or perhaps some rapid assessment of the word's global familiarity; Reichle & Perfetti, 2003) to anticipate how long it will take to identify individual words. As readers become more skilled, they may become increasingly sensitive to these cues and thereby become more adept at initiating saccadic programs so that their eyes leave each word just as it is being identified.

It is also interesting that the agent's solution of using some intermediate stage of lexical processing to initiate saccadic programming is posited to happen in the E-Z Reader model (Pollatsek et al., in press; Rayner et al., 2004; Reichle et al., 1998, 1999, 2003, 2006). In the E-Z Reader model, saccades are initiated prior to word identification—after the completion of an early stage of lexical processing called the *familiarity check*. The present simulations suggest that through learning, readers may be able to tune their word identification and oculomotor systems so that a saccade to move the eyes from one word to the next can be initiated before the fixated word has been completely identified. Doing this will ensure that on average, fixations are neither too short nor too long but are instead just right.

The present simulations thus suggest how a central assumption of the E-Z Reader model—the familiarity check—might come to be instantiated through years of reading experience; that is, through years of practice, readers are able use a variety of cues to initiate saccadic programming prior to lexical access. Our simulations demonstrate that this is possible, thereby enhancing the plausibility of the E-Z Reader model by showing how the assumptions of that model might emerge from fairly simple task constraints and the assumption that readers learn to move their eyes to make reading as efficient as possible. We therefore see the convergence between the present simulations and the E-Z Reader model as increasing the validity of both approaches to understanding eye-movement control during reading.

It is also noteworthy that with the introduction of saccadic error in our second simulation, the adaptive reading agent exhibited two other (unexpected) emergent behaviors. The first was that the simulated landing-site distributions tended to be normally distributed but with more fixations than would be expected by chance landing near the beginnings of the words (see Figure 4). This result is consistent with—and may provide a partial explanation for—the finding that human readers do not fixate the optimal viewing positions most often but instead tend to fixate midway between the beginnings and middles of words (i.e., the preferred viewing location; McConkie et al., 1988, 1991; Rayner, 1979; Rayner et al., 1996). With the agent, this behavior may reflect inherent trade-offs between (a) the goal of maximizing the overall reading rate by making long progressive saccades, (b) the motor error associated with executing saccades (especially those that result in subsequent regressions), and (c) the processing cost associated with trying to identify words from nonoptimal viewing positions. One way to conceptualize these trade-offs is that the agent showed a tendency to play it safe and preferred to make slightly conservative (i.e., shorter) saccades than to make longer saccades followed by frequent regressions. The fact that the so-called words in our simulations were units without sublexical properties (e.g., without differences in the information content of beginning-of-word vs. end-of-word letters, morpheme boundaries, and so forth) suggests that the tendency for readers to fixate the preferred viewing location may not be entirely due to sublexical properties of words. Of course, this conclusion is tentative and must be examined further.

The second unexpected result from Simulation 2 was that the agent exhibited fixation-duration cost for word skipping. It was possible to quantify this finding because it was possible to sort the agent's skips into two groups: those that were deliberate (i.e., those that were part of the policy that was learned by the agent) versus those that were not (i.e., skipping that resulted from saccades overshooting their intended target words). This grouping was informative because only the deliberate skipping of word *n* led to inflated fixation durations on word $n - 1$. This was due to the fact that with deliberate skips, some of the processing of word *n* occurred while the agent's eyes were on word $n - 1$. In contrast, the cost associated with skipping word *n* was always observed on word $n + 1$, irrespective of whether the skips were deliberate or accidental. This latter result stemmed from the fact that whatever processing of word $n + 1$ happened to be completed before the word was fixated had to be done from a distant viewing location (i.e., word $n - 1$) if word *n* was skipped, and because the processing of word *n* often continued from word $n + 1$. Together, these results suggest that the debate about the existence of skipping cost may benefit from the distinction between skipping that is deliberate versus skipping that is not. For example, Kliegl and Engbert (2005) examined a large corpus of eye-tracking data (i.e., the Potsdam corpus; Kliegl et al., 2004) and found that skipping costs on word $n - 1$ are modulated by the frequency and length of word *n*, with costs being evident only if the skipped word was long or infrequent; indeed, they even reported shorter fixation durations prior to short or high-frequency skipped words. Our simulation results suggest that in the Potsdam corpus, readers may have deliberately skipped a larger proportion of the long or infrequent words than the short or frequent words. Future simulations may determine if this interpretation of Kliegl and Engbert's (2005) results is correct.

Finally, we would like to again acknowledge that the present simulations incorporated one assumption that is contentious—that attention is allocated serially during reading. This assumption is central to serial-attention models of eye-movement control (e.g., E-Z Reader; Pollatsek et al., in press; Rayner et al., 2004; Reichle et al., 1998, 1999, 2003, 2006) and was motivated by both findings in the attention literature (Treisman & Souther, 1986; Wheeler & Treisman, 2002) and consideration of the advantage that the serial allocation of attention affords to linguistic processing (Kaiser & Trueswell, 2004; Pollatsek & Rayner, 1999; Vallduvi & Engdahl, 1996). However, we should note that the eye-movement behavior that was ultimately learned by our adaptive reading agent undoubtedly depended upon this assumption. And we should also note that the alternative assumption (i.e., that attention is allocated to more than one word) has also been successfully incorporated into several models of eye-movement control (e.g., Engbert et al., 2002, in press; Kliegl & Engbert, 2003; Reilly, 1993; Reilly & Radach, 2003). It will therefore be important to determine if (and, if so, how) the parallel processing of words would affect the eye-movement behavior of an agent that is faced with the task of

learning to read efficiently. Although the parallel allocation of attention may allow more degrees of freedom in such an agent's behavior (e.g., the cost that results from misplaced fixations may be offset by the fact that several words can be simultaneously processed), the eye-movement behavior that develops in this type of agent may not resemble the eye-movement behavior of real human readers. Future simulations will be necessary to determine if the parallel allocation of attention makes it easier or more difficult for the reading agent to learn to control its eye movement behavior. We therefore suspect that our adaptive reading agent may provide valuable insights into the nature of attention allocation during reading.

References

Altarriba, J., Kroll, J. F., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed language sentences: Evidence from eye fixation and naming times. *Memory & Cognition, 24,* 477–492.

Baird, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. In A. Prieditis & S. J. Russell (Eds.), *Machine learning: Proceedings of the Twelfth International Conference* (pp. 30–37). San Francisco: Morgan Kaufman.

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17,* 364–390.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. Davis, & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge, MA: MIT Press.

Becker, W., & Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. *Vision Research, 19,* 967–983.

Brysbaert, M., & Vitu, F. (1998). Word skipping: Implications for theories of eye movement control in reading. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 125–147). Amsterdam: Elsevier.

Clark, V. P., Fan, S., & Hillyard, S. A. (1995). Identification of early visual evoked potential generators by retinotopic and topographic analyses. *Human Brain Mapping, 2,* 170–187.

Coulom, R. (2002). *Reinforcement learning using neural networks, with applications to motor control.* Unpublished doctoral dissertation, Institut National Polytechnique de Grenoble, Grenoble, France.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20,* 641–655.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research, 42,* 621–636.

Engbert, R., Kliegl, R., & Longtin, A. (2004). Complexity of eye movements in reading. *International Journal of Bifurcation and Chaos, 14,* 493–503.

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112,* 777–813.

Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans: A framework for defining "early" visual processing. *Experimental Brain Research, 142,* 139–150.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14,* 178–210.

Henderson, J. M., & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 417–429.

Hikosaka, O., Takikawa, Y., & Kawagoe, R. (2000). Role of the basal ganglia in the control of purposive saccadic eye movements. *Psychological Review, 80,* 953–978.

Hyönä, J., & Olson, R. K. (1995). Eye movement patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1430–1440.

Inhoff, A. W., Eiter, B. M., & Radach, R. (in press). The time course of linguistic information extraction from consecutive words during eye fixations in reading. *Journal of Experimental Psychology: Human Perception and Performance.*

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics, 40,* 431–439.

Inhoff, A. W., Starr, M., & Shindler, K. L. (2000). Is the processing of words during eye fixations in reading strictly serial? *Perception & Psychophysics, 62,* 1474–1484.

Ishida, T., & Ikeda, M. (1989). Temporal properties of information extraction in reading studied by a text-mask replacement technique. *Journal of the Optical Society A: Optics and Image Sciences, 6,* 1624–1632.

Juhasz, B., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6,* 1312–1318.

Juhasz, B., & Rayner, K. (in press). The role of age-of-acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition.*

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329–354.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension.* Boston: Allyn & Bacon.

Kaiser, E., & Trueswell, J. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition, 94,* 113–147.

Kennedy, A. (1998). The influence of parafoveal words of foveal inspection time: Evidence for a processing trade-off. In G. Underwood (Ed.), *Eye guidance in reading a scene perception* (pp. 149–223). Amsterdam: Elsevier.

Kennedy, A. (2000). Parafoveal processing in word recognition. *Quarterly Journal of Experimental Psychology, 53*A, 429–455.

Kennedy, A., Murray, W. S., & Boissiere, C. (2004). Parafoveal pragmatics revisited. *European Journal of Cognitive Psychology, 16,* 128–153.

Kennedy, A., Pynte, J., & Ducrot, S. (2002). Parafoval-on-foveal interactions in word recognition. *Quarterly Journal of Experimental Psychology, 55*A, 1307–1337.

Kennison, S. M., & Clifton, C. (1995). Determinants of parafoveal preview benefit in high and low working memory capacity readers: Implications for eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 68–81.

Kliegl, R., & Engbert, R. (2003). SWIFT explorations. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 391–411). Oxford, England: Elsevier.

Kliegl, R., & Engbert, R. (2005). Fixation durations before word skipping in reading. *Psychonomic Bulletin & Review, 12,* 132–138.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16,* 262–284.

Klitz, T. S., Legge, G. E., & Tjan, B. S. (2000). Saccade planning in reading with central scotomas: Comparison of human and ideal performance. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (pp. 667–682). Oxford, England: Elsevier.

Lee, H.-W., Legge, G. E., & Ortiz, A. (2003). Is word recognition different in central and peripheral vision? *Vision Research, 43,* 2837–2846.

Leff, A. P., Scott, S. K., Rothwell, J. C., & Wise, R. J. S. (2001). The

planning and guiding of reading saccades: A repetitive transcranial magnetic stimulation study. *Cerebral Cortex, 11,* 918–923.

Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., & Tjan, B. S. (2002). Mr. Chips 2002: New insights for an ideal-observer model of reading. *Vision Research, 42,* 2219–2234.

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: An ideal-observer model of reading. *Psychological Review, 104,* 524–553.

McConkie, G. W., Kerr, P. W., & Dyre, B. P. (1994). What are "normal" eye movements during reading: Toward a mathematical description. In J. Ygge & G. Lennerstrand (Eds.), *Eye movements in reading* (pp. 315–328). New York: Pergamon Press.

McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of initial eye fixations in words. *Vision Research, 28,* 1107–1118.

McConkie, G. W., Zola, D., Grimes, J., Kerr, P. W., Bryant, N. R., & Wolff, P. M. (1991). Children's eye movements during reading. In J. F. Stein (Ed.), *Vision and visual dyslexia 13.* London: MacMillan.

McDonald, S. A., Carpenter, R. H. S., & Shillcock, R. C. (2005). An anatomically constrained, stochastic model of eye movement control in reading. *Psychological Review, 112,* 814–840.

McPeek, R. M., Skavenski, A. A., & Nakayama, K. (2000). Concurrent processing of saccades in visual search. *Vision Research, 40,* 2499–2516.

Molker, A., & Fischer, B. (1999). The recognition and correction of involuntary prosaccades in an antisaccade task. *Experimental Brain Research, 125,* 511–516.

Mouchetant-Rostaing, Y., Giard, M.-H., Bentin, S., Aguera, P.-E., & Pernier, J. (2000). Neurophysiological correlates of face gender processing in humans. *European Journal of Neuroscience, 12,* 303–310.

Murray, W. S. (1998). Parafoval pragmatics. In G. Underwood (Ed.), *Eye guidance in reading a scene perception* (pp. 181–199). Amsterdam: Elsevier.

Nuthmann, A., Engbert, R., & Kliegl, R. (2005). Mislocated fixations during reading and the inverted optimal viewing position effect. *Vision Research, 45,* 2201–2217.

O'Regan, J. K. (1990). Eye movements and reading. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes* (pp. 395–453). Amsterdam: Elsevier.

O'Regan, J. K. (1992). Optimal viewing position in words and the strategy-tactics theory of eye movements in reading. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 333–354). Springer-Verlag.

O'Regan, J. K., & Lévy-Schoen, A. (1987). Eye movement strategy and tactics in word recognition and reading. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 363–383). Hillsdale, NJ: Erlbaum.

O'Regan, J. K., Lévy-Schoen, A., Pynte, J., & Brugaillère, B. (1984). Convenient fixation location within isolated words of different length and structure. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 250–257.

Perfetti, C. A. (1985). *Reading ability.* New York: Oxford University Press.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). Philadelphia: John Benjamins.

Pollatsek, A., & Rayner, K. (1999). Is covert attention really unnecessary? *Behavioral and Brain Sciences, 22,* 695–696.

Pollatsek, A., Rayner, K., & Balota, D. A. (1986). Inferences about eye movement control from the perceptual span in reading. *Perception & Psychophysics, 40,* 123–130.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology, 52,* 1–56.

Pynte, J., Kennedy, A., & Ducrot, S. (2004). The influence of parafoveal typographical errors on eye movements in reading. *European Journal of Cognitive Psychology, 16,* 178–202.

Radach, R., & Heller, D. (2000). Relations between spatial and temporal aspects of eye movement control. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (pp. 165–192). Oxford, England: Elsevier.

Rayner, K. (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition, 4,* 443–448.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85,* 618–660.

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception, 8,* 21–30.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372–422.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 720–732.

Rayner, K., & Bertera, J. H. (1979). Reading without a fovea. *Science, 206,* 468–469.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14,* 191–201.

Rayner, K., & Juhasz, B. (2004). Eye movements in reading: Old questions and new directions. *European Journal of Cognitive Psychology, 16,* 340–352.

Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements? *Vision Research, 16,* 829–837.

Rayner, K., & Morrison, R. M. (1981). Eye movements and identifying words in parafoveal vision. *Bulletin of the Psychonomic Society, 17,* 135–138.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading.* Englewood Cliffs, NJ: Prentice Hall.

Rayner, K., Pollatsek, A., & Reichle, E. D. (2003). Eye movements in reading: Models and data. *Brain and Behavioral Sciences, 26,* 507–526.

Rayner, K., & Raney, G. E. (1996). Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review, 3,* 238–244.

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: A comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1188–1200.

Rayner, K., Slowiaczek, M. L., Clifton, C., & Bertera, J. H. (1983). Latency of sequential eye movements: Implications for reading. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 912–922.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint of eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3,* 504–509.

Rayner, K., White, S. J., Kambe, G., Miller, B., & Liversedge, S. P. (2003). On the processing of meaning from parafoveal vision during eye fixations in reading. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 213–234). Oxford, England: Elsevier.

Reichle, E. D., & Laurent, P. A. (2004, September). *The emergence of "intelligent" eye-movement control during reading: A computational account*. Poster session presented at Architectures and Mechanisms for Language Processing 2004, Aix-en-Provence, France.

Reichle, E. D., & Perfetti, C. A. (2003). Morphology in word identification: A word experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading, 7,* 219–237.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Towards a model of eye movement control in reading. *Psychological Review, 105,* 125–157.

Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the E-Z Reader model. *Vision Research, 39,* 4403–4411.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparison to other models. *Brain and Behavioral Sciences, 26,* 445–476.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement control during reading. *Cognitive Systems Research, 7,* 4–22.

Reilly, R. (1993). A connectionist framework for model eye movement control in reading. In G. d'Ydewalle & J. Van Rensbergen (Eds.), *Perception and cognition: Advances in eye movement research* (pp. 193–212). Amsterdam: North-Holland.

Reilly, R., & O'Regan, J. K. (1998). Eye movement control in reading: A simulation of some word-targeting strategies. *Vision Research, 38,* 303–317.

Reilly, R., & Radach, R. (2003). Foundations of an interactive activation model of eye movement control in reading. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 429–455). Oxford, England: Elsevier.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Salvucci, D. D. (2000). An integrated model of eye movements and visual encoding. *Cognitive Systems Research, 1,* 201–220.

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition, 26,* 1270–1281.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80,* 1–27.

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

Sereno, S. C. (1992). Early lexical effects when fixating a word in reading. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 304–316). New York: Springer-Verlag.

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *NeuroReport, 9,* 2195–3000.

Thibadeau, R., Just, M. A., & Carpenter, P. A. (1982). A model of the time course and content of reading. *Cognitive Science, 6,* 157–203.

Treisman, A., & Souther, J. (1986). Illusory words: The roles of attention and of top-down constraints in conjoining letters to form words. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 3–17.

Vallduvi, E., & Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics, 34,* 459–519.

Van Rullen, R., & Thorpe, S. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13,* 454–461.

Vergilino, D., & Beauvillain, C. (2000). The planning of refixation saccades in reading. *Vision Research, 40,* 3527–3538.

Vitu, F., Brysbaert, M., & Lancelin, D. (2004). A test of parafoveal-on-foveal effects with pairs of orthographically related words. *European Journal of Cognitive Psychology, 16,* 154–177.

Vitu, F., McConkie, G. W., Kerr, P., & O'Regan, J. K. (2001). Fixation location effects on fixation locations during reading: An inverted optimal viewing position effect. *Vision Research, 41,* 3513–3533.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General, 131,* 48–64.

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1,* 202–238.

Wolfe, J. M., & Bennett, S. C. (1996). Preattentive object files: Shapeless bundles of basic features. *Vision Research, 37,* 25–43.

Wolverton, G. S., & Zola, D. (1983). The temporal characteristics of visual information extraction during reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes* (pp. 41–52). New York: Academic Press.

Wörgötter, F. (2004, February–March). *Actor-critic model of animal control: A critique of reinforcement learning.* Paper presented at the Fourth International ICSC Symposium on Engineering of Intelligent Systems, Madeira, Portugal.

Yang, S.-N., & McConkie, G. W. (2001). Eye movements during reading: A theory of saccade initiation time. *Vision Research, 41,* 3567–3568.

Yang, S.-N., & McConkie, G. W. (2004). Saccade generation during reading: Are words necessary? *European Journal of Cognitive Psychology, 16,* 226–261.

Received March 2, 2005
Revision received September 14, 2005
Accepted September 15, 2005 ■

Neural Networks 21 (2008) 1493–1499

Contents lists available at ScienceDirect

Neural Networks

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

The emergence of saliency and novelty responses from Reinforcement Learning principles☆

Patryk A. Laurent*

University of Pittsburgh, Centers for Neuroscience and for the Neural Basis of Cognition, 623 LRDC, 3939 O'Hara St., Pittsburgh, PA 15260, USA

ARTICLE INFO

Article history:
Received 1 October 2007
Received in revised form
15 September 2008
Accepted 18 September 2008

Keywords:
Novelty response
Reinforcement learning
Dopamine
Orienting
Reward-prediction error

ABSTRACT

Recent attempts to map reward-based learning models, like Reinforcement Learning [Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An introduction*. Cambridge, MA: MIT Press], to the brain are based on the observation that phasic increases and decreases in the spiking of dopamine-releasing neurons signal differences between predicted and received reward [Gillies, A., & Arbuthnott, G. (2000). Computational models of the basal ganglia. *Movement Disorders*, *15*(5), 762–770; Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27]. However, this *reward-prediction error* is only one of several signals communicated by that phasic activity; another involves an increase in dopaminergic spiking, reflecting the appearance of salient but unpredicted non-reward stimuli [Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4–6), 495–506; Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, *96*(4), 651–656; Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews Neuroscience*, *7*(12), 967–975], especially when an organism subsequently orients towards the stimulus [Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27]. To explain these findings, Kakade and Dayan [Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, *15*(4–6), 549–559.] and others have posited that novel, unexpected stimuli are intrinsically rewarding. The simulation reported in this article demonstrates that this assumption is not necessary because the effect it is intended to capture emerges from the reward-prediction learning mechanisms of Reinforcement Learning. Thus, Reinforcement Learning principles can be used to understand not just reward-related activity of the dopaminergic neurons of the basal ganglia, but also some of their apparently non-reward-related activity.

Reinforcement Learning (*RL*) is becoming increasingly important in the development of computational models of reward-based learning in the brain (Gillies & Arbuthnott, 2000). RL is a class of computational algorithms that specifies how an artificial "agent" (e.g., a real or simulated robot) can learn to select actions in order to maximize total expected reward (Sutton & Barto, 1998). In these algorithms, an agent bases its actions on *values* that it learns to associate with various *states* (e.g., the perceptual cues associated with a stimulus). These values can be gradually learned through *temporal-difference learning*, which adjusts state values based on the difference between the agent's existing reward prediction for the state and the actual reward that is subsequently obtained from the environment. This computed difference, termed *reward-prediction error*, has been shown to correlate very well with the phasic activity of dopamine-releasing neurons projecting from the substantia nigra in non-human primates (Schultz, 1998). Furthermore, in humans, the striatum, which is an important target of dopamine, exhibits an fMRI BOLD signal that appears to reflect reward-prediction error during reward-learning tasks (McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Tanaka et al., 2004). This fMRI finding complements the physiology data because striatal BOLD is assumed to reflect, at least in part, afferent synaptic activity (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001) and the dopamine neurons project heavily to the striatum.

Although the aforementioned physiological responses appear to be related to the reward-prediction computations of RL, there is also an increase in dopaminergic phasic activity in response to arousing and/or novel stimuli that is seemingly unrelated to reward (Dommett et al., 2005; Doya, 2002; Horvitz, 2000; Redgrave, Prescott, & Gurney, 1999). A similar phenomenon has been recently observed in humans using fMRI (Bunzeck & Düzel, 2006). There are several reasons why this "novelty" or "saliency" response is said to be unrelated to reward-prediction error: (1) it appears very early,

☆ Contributed article.

* Tel.: +1 412 624 3191; fax: +1 412 624 9149.
E-mail address: patryk@cnbc.cmu.edu.

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.09.004

before the identity of the stimulus has been assessed, so that an accurate reward prediction cannot be generated; (2) it corresponds to an *increase* in neural activity (i.e., it is positive) for both aversive and appetitive stimuli; and (3) it habituates (Redgrave & Gurney, 2006). Indeed, these saliency/novelty responses of the dopamine-releasing neurons are most reliable when the stimuli are unpredicted and result in orienting and/or approach behavior (Schultz, 1998) regardless of the eventual outcome, highlighting the fact that they are qualitatively different from learned reward prediction. The challenge, therefore, has been to explain this apparent paradox (i.e., how novelty affects the reward-prediction error) within the theoretical framework of RL.

Kakade and Dayan (2002) attempted to do exactly this; in their article, they postulate two ways in which novelty responses could be incorporated into RL models of dopaminergic function—both involved the inclusion of new theoretical assumptions. The first assumption, referred to as *novelty bonuses*, involves introducing an additional reward when novel stimuli are present, above and beyond the usual reward received by the agent. This additional reward enters into the computation so that learning is based on the difference between the agent's existing reward prediction and the sum of both the usual reward from the environment and the novelty bonus. Thus, the novelty becomes part of the reward that the agent is attempting to maximize. The second assumption, termed *shaping bonuses*, can be implemented by artificially increasing the values of states associated with novel stimuli. Because the temporal-difference learning rule used in RL is based on the difference in reward-prediction between successive states, the addition of a constant shaping bonus to states concerned with the novel stimuli has no effect on the final behavior of the agent. However, a novelty response still emerges when the agent enters the part of the state space that has been "shaped" (i.e., that is associated with novelty).

Although the addition of each of these assumptions is sufficient to explain many observed effects of novelty, the assumptions also interfere with the progression of learning. As Kakade and Dayan (2002) point out, novelty bonuses can distort the value function (i.e., the values associated with each state by the agent) and affect what is ultimately learned because they are implemented as an additional reward that is intrinsically associated with novel states. The problem is that the agent learns to predict both the primary and novelty components of the reward. Although Kakade and Dayan point out that shaping bonuses do not cause this type of problem because they become incorporated into the reward predictions from preceding states, their addition is still problematic because shaping bonuses introduce biases into the way an agent will explore its state space. Thus, although these additional assumptions may explain how novelty affects the reward-prediction error in RL, they are problematic. Further, the explanations come at the cost of reducing the parsimony of modeling work that attempts to use RL to understand the behavior of real biological organisms.

The simulation reported below was carried out in order to test the hypothesis that a simple RL agent, without any additional assumptions, would develop a reward-prediction error response that is similar to the non-reward-related dopamine responses that are observed in biological organisms. An RL agent was given the task of interacting with two types of object – one positive and the other negative – that appeared at random locations in its environment. In order to maximize its reward, the agent had to learn to approach and "consume" the positive object, and to avoid (i.e., not "consume") the negative object. There were three main predictions for the simulation.

The first prediction was simply that, in order to maximize its reward, the agent would in fact learn to approach and "consume" the positive, rewarding objects while simultaneously learning to avoid the negative, punishing objects. The second prediction was slightly less obvious: that the agent would exhibit an orienting response (i.e., learn to shift its orientation) towards both negative and positive objects. This prediction was made because although the agent could "sense" the appearance of an object and its location, the positive or negative identity of the object (i.e., the cue that the agent would eventually learn to associate with the reward value of the object) could not be determined by the agent until *after* the agent had actually oriented towards the object. Finally, the third (and most important) prediction was related to the simulated dopaminergic phasic response in the model; this prediction was that, when the object appeared, the agent would exhibit a reward-prediction error that was computationally analogous to the phasic dopamine response observed in biological organisms, being positive for both positive *and* negative objects. This response was also predicted to vary as a function of the distance between the agent and the stimulus, which in the context of the simulation was a proxy measure for stimulus "intensity" or salience. As will be demonstrated below, these predictions were confirmed by the simulation results, demonstrating that the apparently non-reward-related dopamine responses can in principle emerge from the basic principles of RL. The theoretical implications of these results for using RL to understand non-reward-related activity in biological organisms will be discussed in the final section of this article.

1. Method

As already mentioned, RL algorithms specify how an agent can use moment-to-moment numerical rewards to learn which actions it should take in order to maximize the total amount of reward that it receives. In most formulations, this learning is achieved by using reward-prediction errors (i.e., the difference between an agent's current reward prediction and the actual reward that is obtained) to update the agent's reward predictions. As the reward predictions are learned, the predictions can also be used by an agent to select its next action. The usual *policy* (defined in Eq. (2)) is for the agent to select the action that is predicted to result in the largest reward. The actual reward that is provided to the agent at any given time is the sum of the immediate reward plus some portion of the value of the state that the agent enters when the action is completed. Thus, if the agent eventually experiences positive rewards after having been in a particular state, the agent will select actions in the future that are likely to result in those rewarded states; conversely, if the agent experiences negative rewards (i.e., punishment) it will avoid actions in the future that lead to those "punished" states.

The specific algorithm that determines the reward predictions that are learned for the various states (i.e., the *value function* V) is called *Value Iteration*[1] and can be formally described as:

[1] Another Reinforcement Learning algorithm, called *Trajectory Sampling* (Sutton & Barto, 1998), is frequently used instead of Value Iteration when the state space becomes so large that it cannot be exhaustively iterated or easily stored in a computer's memory. Rather than iterating over every state in the state space and applying the value function update equation based on the actions that appear to lead to the most reward, Trajectory Sampling works by following paths through the state space. Similar to Value Iteration, the actions leading to the most reward are usually selected from each state, but occasionally a random exploratory action is chosen with some small probability. Thus the algorithm is: *From some starting states, select an action leading to the most reward [e.g.,* reward $+ \gamma V(s')$*] with probability* ε*, or select a random exploratory action with probability* $1 - \varepsilon$*. Apply* $V(s) \leftarrow V(s) + \alpha$*[*reward $+ \gamma V(s') - V(s)$*] during non-exploratory actions from states.*

Besides overcoming the technical limitations of computational time and memory, Trajectory Sampling may be appealing because it may better reflect the manner in which real biological organisms learn: by exploring paths in a state space. On the task described in this paper, Trajectory Sampling yields results that are qualitatively identical to those obtained with Value Iteration. However, for conciseness those

For all possible states s,

$$V(s) \leftarrow V(s) + \alpha \left[\max_{\text{action}\in M} \{\text{reward} + \gamma V(s')\} - V(s) \right] \quad (1)$$

where s corresponds to the current state, $V(s)$ is the current reward prediction for state s that has been learned by the agent, $\max_{\text{action}\in M}\{\}$ is an operator for the maximum value of the bracketed quantity over the set of all actions M available to the agent, $V(s')$ is the agent's current reward prediction for the next state s', α is some learning rate (between 0 and 1), and γ is a discount factor reflecting how future rewards are to be weighted relative to immediate rewards. The initial value function was set so that $V(s)$ was 0 for all states s.

The value function $V(s)$ was implemented as a lookup table, which is formally equivalent to the assumption of perfect memory. Although function approximators such as neural networks have been used with some success to represent value functions (Baird, 1995), a lookup table was used to ensure that the results were not dependent on the types of generalization mechanism that are provided by various function approximators. The agent was trained for 1500 learning iterations over its state space. Because of the unpredictability of the identity of the objects, a value function update parameter of less than one ($\alpha = 0.01$) was used during the learning to allow for averaging of different outcomes. Finally, the discount factor was set to $\gamma = 0.99$ to encourage the agent to seek reward sooner rather than delay its approach behavior until the end of the trial (although changing it from a default value of 1 had no effect on the results reported here). In order to independently determine whether 1500 learning iterations were sufficient for learning to complete, the average amount of change in the learned was monitored and was found to have converged before this number of iterations.

After training, the specific algorithm that governs the agent's behavior (i.e., the *policy* of actions that it takes from each given state) is:

$$\pi(s) = \underset{\text{action}\in M}{\text{argmax}} [\text{reinforcement} + \gamma V(s')] \quad (2)$$

where $\pi(s)$ is the action the agent will select from state s, and the right side of the equation returns the action (e.g., change of orientation, movement, or no action) which maximizes the sum of the reward and the discounted value of the resulting state s'.

In the simulation that is reported below, all of the states that were visited by the agent were encoded as 7-dimensional vectors that represented information about both the external "physical" state of the agent and its internal "knowledge" state. The physical information included both the agent's current position in space and its orientation. The knowledge information included the position of the object (if one was present) and the identity of that object (if it had been determined by the agent). The specific types of information that were represented by the agent are shown in Table 1.

There were a total of 21,120 states in the simulation[2]. However, the states in which there was an unidentified positive and unidentified negative object are, from the perspective of the agent, identical, so there are therefore only 16,280 distinct states. Thus, during each iteration of learning, it was necessary to visit some of those "identical" states twice to allow for the fact that half of the time they might be followed with the discovery of a positive object, and half of the time they might be followed with the discovery of a negative object.[3]

At the beginning of each simulated testing trial, the agent was placed in the center of a simulated linear 11 × 1 unit track with five spaces to the "east" (i.e., to the right) of the agent and five spaces to the "west" (i.e., to the left) of the agent. As Table 1 shows, the agent's state-vector included an element indicating its current location on the track (i.e., an integer from 0 to 10), as well as an element (i.e., a character "*n*", "*s*", "*e*", or "*w*") representing its current orientation (i.e., north, south, east, or west, respectively). The agent's initial orientation was always set to be "north," and no other object was present in the environment (i.e., the value of "OBJECT" in the agent's state-vector was set to equal to "0").

During each time-step of the simulation, the agent could perform one of the following actions: (1) do nothing, and remain in the current location and orientation; (2) orient to the north, south, east or west; or (3) move one space in the environment (east or west). The result of each action took place on the subsequent simulated time-step. All changes in the location and/or orientation of the agent in space occurred through the selection of actions by the agent. However, during every time-step of the simulation, even when a "do nothing" action was selected, time was incremented by 1 until the end of the trial (i.e., time-step 20).

The agent's environment was set up so that half of the time, an object appeared at a random location (but not in the same location as the agent) after ten time steps; 50% of the objects were positive (represented by a "+"; see Table 1) and 50% of the objects were negative (represented by a "−"). The delay before the object appeared was introduced to allow the observation of any behavior the agent may have exhibited before the appearance of the object. If the agent was not oriented towards the object when it appeared, then the element representing the "OBJECT" identity in the agent's state vector was changed from "0" to "?" to reflect the fact that the identity of the object that was now present was currently unknown. However, if the agent was oriented towards the object, then on the subsequent time-step the "OBJECT" element was set to equal to the identity of the object, so that "0" became either "+" or "−" for positive and negative objects, respectively.

If the agent moved to an object's location, then during the next time-step the object vanished. If the object had been positive, then the agent's "CONSUMED" flag was set equal to true and the agent was rewarded (reward = +10); however, if the object

results are not reported here in detail. Value Iteration was selected for the simulation in this paper for two main reasons. First, because Trajectory Sampling involves stochasticity in the selection of trajectories; the large amount of branching that is due to the many possible sequences of actions in this task may result in agents that lack experience with some states unless the exploration–exploitation parameter (i.e., ε-greediness (Sutton & Barto, 1998)) is carefully selected. This lack of experience with particular states can be disruptive of an agent's performance when a lookup table memory structure is used because of the lack of generalization of value to similar (but possibly unvisited) states. Thus, it was preferred to take advantage of the exhaustive exploration of state space that is guaranteed with Value Iteration. Second, the use of Value Iteration obviated the need to specify that additional exploration–exploitation parameter, thereby simplifying the simulation. Note that Trajectory Sampling can ultimately approximate Value Iteration as the number of trajectories approaches infinity (Sutton & Barto, 1998).

[2] The number of 21,120 states can be calculated as follows: 11 possible agent locations × 4 possible agent orientations × (10 time-steps before an object might appear + 10 time-steps where no object appeared + 10 time-steps where the agent had been positively reinforced + 10 time-steps where the object had been negatively reinforced + 11 possible object locations × (10 time-steps with a positive identified object + 10 time-steps with a negative identified object + 10 time-steps with an unidentified positive object + 10 time-steps with an unidentified negative object))].

[3] The existence of these "hidden" states must be considered during training because Value Iteration only looks "one step ahead" from each state in the state space. The fact that states with negative and positive unidentified objects are effectively identical would prevent learning about and averaging the values in the two different subsequent states in which either the positive or negative object becomes identified. A Trajectory Sampling approach on the other hand maintains the hidden state information (i.e., the identity of the unidentified stimulus) throughout the trial and so with that variant of RL the hidden states are not a concern.

Table 1
The dimensions used in the RL simulations and the possible values of those dimensions.

Dimension no.	Description	Possible values
1	Position of agent	Integer (0–10)
2	Orientation of agent	Character ("*n*", "*s*", "*e*", or "*w*")
3	Position of object	Integer (0–10)
4	Identity of object	Character ('0", "?", "+", or "−")
5	"Shocked" by object	Boolean (true or false)
6	"Consumed" object	Boolean (true or false)
7	Time since trial onset	Integer (0–20)

had been negative, then the "SHOCKED" flag was set to true and the agent was punished (reward $= -10$). (Note that the flags were set in this way regardless of whether the agent had or had not identified the object; e.g., the agent could consume an object without ever orienting towards it.) On the subsequent time-step, the "SHOCKED" or "CONSUMED" flag was cleared. The agent was also given a small penalty (reinforcement $= -1$) for each movement or orienting action, and received no reward or punishment (reinforcement $= 0$) if it performed no action.

Both the overt behaviors (i.e., orienting and movement) and a measure of reward-prediction error were quantified for the agent. The overt behavior (i.e., the list of actions selected by the agent) was used as an indication of whether the task had been learned. The measure of reward-prediction error was used to test the hypothesis about the emergence of the non-reward dopaminergic phasic signal. The reward-prediction error, δ, was measured at the time t of the appearance of an object by subtracting the reward prediction at the previous time-step, i.e., $V(s)$ at time step $t-1$, from the reward prediction when the object appeared, i.e., $V(s)$ at time t, yielding the quantity $\delta = V(s_t) - V(s_{t-1})$.

2. Results

Simulated behavior. The overt behavior of the agents was first quantified. The results of this analysis showed that, after training, the agent approached and obtained positive reinforcement from all of the positive objects and never approached any of the negative objects. Together, these results provide behavioral confirmation that the agents learned to perform the task correctly. This conclusion is bolstered by the additional observation that, during the trials when no object appeared, the agent remained motionless. As predicted, the agent oriented to both positive and negative objects.

Simulated reward-prediction error. The central hypothesis of this paper is that the appearance of an unpredictable stimulus will consistently generate a positive reward-prediction error, even if that object happens to be a "negative" object that is always punishing. In support of this hypothesis, the agent exhibited a positive reward-prediction error whenever an (unidentified) object appeared, but not when nothing appeared. Also consistent with the central hypothesis is the fact that the magnitude of the agent's phasic response (δ, measured as described in the Method section) was sensitive to the simulated "intensity" of the stimulus, defined using the distance between the agent and the object (see Fig. 1). A regression analysis indicated that the magnitude of δ was inversely related to the distance from the object, so that closer objects caused a stronger response ($r = -0.999$, $p < 0.001$; $\beta = 0.82$). This negative correlation was caused by the small penalty (reinforcement $= -1$) that was imposed for each movement that the agent was required to make in order to approach the positive object, consume it, and thereby obtain reward.

Given that positive and negative objects appeared in this simulation with equal probability ($p = .25$), the question arises: Why was the agent's reward-prediction error signal positive at the time of the object's appearance? Reasoning along the lines of Kakade and Dayan (2002), one might predict that the signal should reflect the average of all of the learned rewards from such situations, and therefore be equal to zero. The key to understanding this result is to note that not only does RL make an agent less likely to choose actions that result in negative reinforcement, it also makes an agent less likely to enter states which eventually lead to negative reinforcement. This results in a kind of "higher-order" form of learning that is depicted in Fig. 2 and described next.

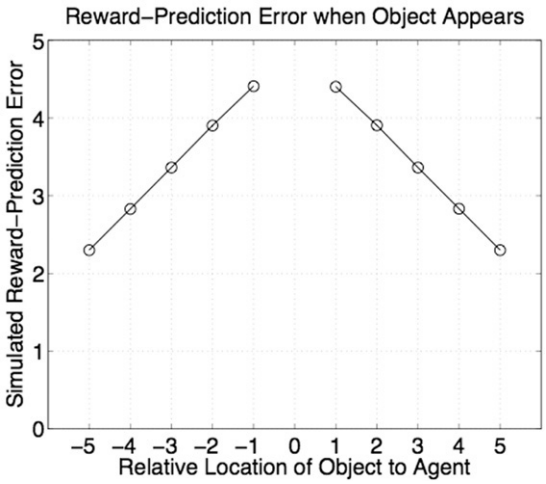

Fig. 1. This figure shows the reward-prediction error (i.e., δ) when the object appeared as a function of the location of the object relative to the location of the agent. The responses are identical for both positive and negative objects. When no object appeared, the response was 0. Note that the size of the response is inversely correlated with distance from the object when it appeared. There is no data for location 0 because the object would be immediately consumed had it appeared there.

At the beginning of learning (see Fig. 2A), the agent orients to both "+" and "−" objects, approaches them, and is both rewarded and punished by consuming each type of object. If the agent's learned state values were unable to influence the agent's actions (see Fig. 2B), then the agent would continue to approach and consume the objects. The appearance of the cue would then predict an average reward of 0 and there would not be a sudden increase in reward-prediction error. However, the agent in this simulation *does* use learned state values to influence its actions (see Fig. 2C), and although the agent still has to orient to the unknown object to determine its identity, it will no longer consume a negative object if it approached it (as it might if trained with a random exploration algorithm like *trajectory sampling* (see Footnote 1)). Furthermore, because temporal-difference learning allows the negative reward prediction to "propagate" back to preceding states, and because there is a small cost for moving in space, the agent learns to avoid approaching the negative object entirely. Thus, after this information has been learned, the value of the state when the object first appears (indicated as "V" in the first circle in each sequence) is not based on the average of the positive and negative outcome state values, but is instead based on the average of positive and the "neutral" outcome that is attained once the agent learns to avoid the negative objects. This is why the

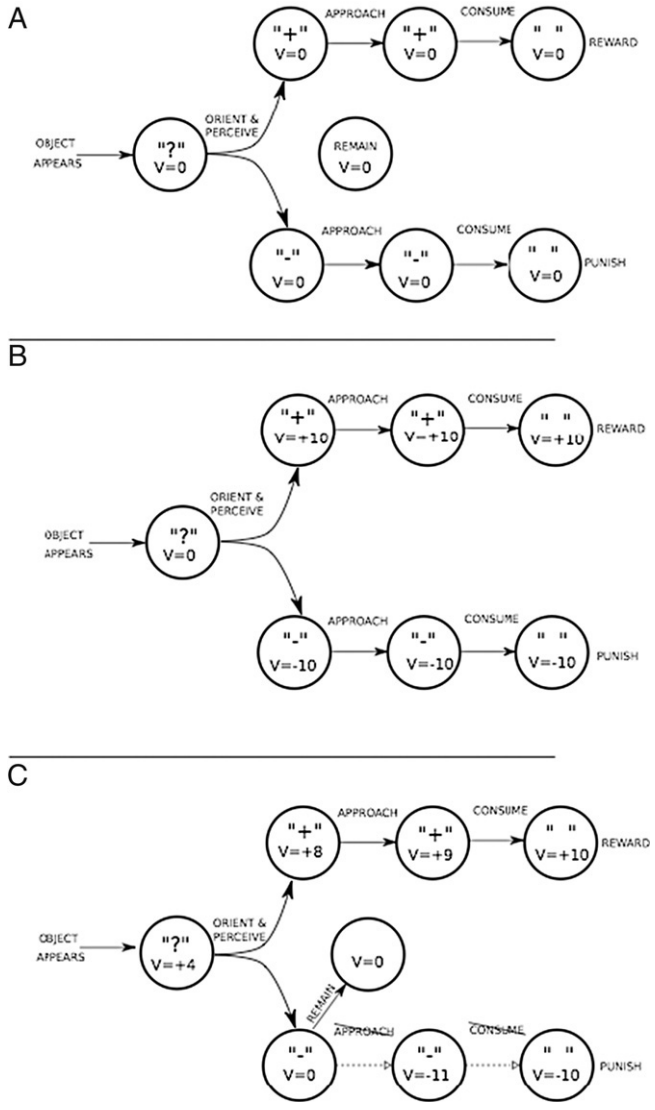

Fig. 2. Illustration showing how an RL agent develops positive reward-prediction error when it is trained with both rewarding and punishing stimuli in its environment and is able to choose whether to approach and consume them. (A) The situation before learning: all states begin with a value of 0, and the agent has not yet learned the rewarding and punishing values of the "+" and "−" stimuli. (B) A temporal-difference learning algorithm is used without allowing those values to affect the actions of the agent: the agent learns reward predictions based on experience but is unable to use the learned values to influence its own behavior. In this case, the reward-prediction error when the object appears will be the average of the positive and negative outcomes (i.e., 0). (C) We show what happened in the present simulation. The agent quickly learns to avoid consuming, or even approaching, the negative object. The result is that when the stimulus appears, the reward-prediction error is based on the average of the positive stimulus and a neutral outcome in which the negative stimulus is avoided and is consistently greater than 0. Note: This figure does not illustrate the fact that in the present simulation, more distant objects require more actions (and therefore more intervening small punishments) in order to approach them. That fact is what causes the decreasing magnitude of the novelty/saliency response for objects that appear more distantly from the agent (e.g., as plotted in Fig. 2).

average of all rewards actually obtained by the trained agent was greater than zero, and explains why the agent's reward prediction (and therefore reward-prediction error when the object suddenly appears) was a net positive. This is illustrated in Fig. 3. In fact, as long as the agent can learn to change its behavior and avoid the negative object, the value of the negative object is ultimately irrelevant to the final behavior of the agent and the magnitude of the novelty/saliency response.

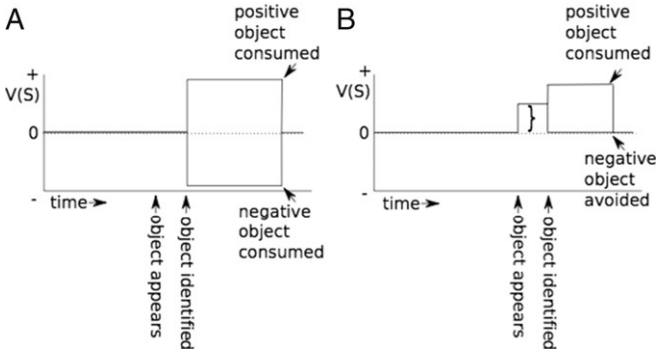

Fig. 3. (A) Demonstrates the changes in reward prediction that would have occurred if RL did not result in higher-order learning (i.e., if the agent could not take measures to avoid the negative outcome), so that the agent was forced to consume all the objects that appeared. When an object appears, the agent does not know yet its identity but generates a net reward prediction of zero because the reward prediction is the average of the positive and negative consequences (i.e., half the time the object has been positive, and half the time it has been negative). (B) Demonstrates what actually occurred: higher-order learning permitted the agent to avoid the negative object, so that when the stimulus appeared, the agent had a greater-than-0 reward prediction because it is the average of the positive outcome and null outcomes. The curly brace spans the difference in reward-prediction values that represents this reward-prediction error.

The simulation results are critically dependant on three assumptions. First, the stimuli had to be "salient" in that the magnitude of the reinforcement predicted by the initial cue was sufficiently large (e.g., +10) relative to the costs of orienting and approaching (e.g., −1). If the magnitude had been relatively small, the agent would not have learned to orient, nor would it have generated the positive reward-prediction error response. Second, a delay prior to recognizing the stimuli was also necessary. (Delay is a proxy for "novelty" under the reasoning that a familiar stimulus would be quickly recognized.) Without a delay, the agent would have simply generated the appropriate positive or negative reward prediction error for the actual perceived object. Finally, the agent's behavior had to be determined by the values that it had learned. If the agent could not use learned values to control its own behavior (i.e., whether to approach the stimuli), then its reward prediction when an object appeared would have equaled 0, the average of the equiprobable positive and negative outcomes.

3. General discussion

The simulation reported in this article demonstrated that a positive reward-prediction error occurs when an unpredictable stimulus, either rewarding or punishing, appears but cannot be immediately identified. Furthermore, the simulation indicated that the size of the reward-prediction error increases with proximity of the stimulus to the agent, which in the context of the simulation is a proxy for stimulus intensity and is thus related to salience. In the theoretical framework of RL, reward predictions are normally understood to reflect the learned value of recognized stimuli, or of the physical and/or cognitive states of an agent (Reichle & Laurent, 2006). However, the reward-prediction error reported here has a qualitatively different interpretation because it is generated before the agent has recognized the object. Together, these results support the hypothesis that RL principles are sufficient to produce a response that is seemingly unrelated to reward, but instead related to the properties of novelty and saliency. This conclusion has several important ramifications for our general understanding of RL and for our interpretation of RL as an account of reward learning in real biological organisms.

First, the reward prediction that is generated by an RL agent when an unidentified stimulus appears is not necessarily a strict average of the obtainable rewards as suggested by Kakade and

Dayan (2002), but can in fact be greater in magnitude than that particular average. Kakade and Dayan would predict that the average reward prediction should be equal to zero because, the trials were rewarded and punished equally often. This surprising result emerged because the agent learned in an "on-policy" manner; that is, the agent learned not only about negative outcomes, but also about its ability to avoid those outcomes. This ability of the reward system to cause an agent to avoid negative outcomes should be carefully considered in translating our understanding of RL to real organisms. This fact is potentially even more important given the apparent asymmetry in the capacity of the dopaminergic phasic response to represent positive reward prediction error better than negative reward prediction error (Niv, Duff, & Dayan, 2005). It may be sufficient to indicate that a particular sequence of events leads to a negative outcome, but that for the purposes of action selection, the magnitude of that outcome is unimportant.

A second ramification of the current simulation is that the novelty response may emerge from an interaction between perceptual processing systems and reward-prediction systems. Specifically, the novelty response may be due to a form of similarity between novel objects and objects that have not yet undergone complete perceptual processing[4]. In this simulation, novelty was implemented by introducing a delay before the object's identity (and consequently its rewarding or punishing nature) became apparent to the agent. This was done under the assumption that novel objects take longer to identify, but this assumption also resulted in the positive and negative objects being perceived similarly when they first appeared (i.e., they were both encoded as "?"). In contrast, Kakade and Dayan (2002) suggest that novelty responses and "generalization" responses are essentially different despite being manifested similarly in the neurophysiology data.

A third ramification of the current simulation results is that they show that the additional assumptions of novelty and shaping bonuses that were proposed by Kakade and Dayan (2002) are not necessary. Instead, novelty-like responses can emerge from realistic perceptual processing limitations and the knowledge of being able to avoid negative outcomes. This is fortunate because, as pointed out by Kakade and Dayan, novelty bonuses distort the value function that is learned by an agent, and shaping bonuses affect the way in which agents explore their state spaces. The inclusion of either of these assumptions thus reduces the parsimony of models based on RL theory. Interestingly, the results presented here also help explain why the biological novelty response might not be disruptive to reward-based learning in real organisms: the novelty response is in fact already predicted by RL. That is, the novelty response reflects behaviors and reward predictions that are inherent in an agent that has already learned something about its environment.

An alternative (but not mutually exclusive) interpretation of the present simulation results is that there is indeed an abstract (perhaps cognitive) reward that agents obtain by orienting towards and identifying objects. In studies of dopaminergic activity, positive phasic responses can occur to unanticipated cues that are known to predict a reward. This simulation, however, demonstrates how these kinds of responses can also occur in response to a cue that could ultimately predict either reward or punishment. The only consistent benefit that is predicted by the cue is the gain in information obtained when the agent determines the identity of the object. Thus, if there is a valid, learned "reward prediction" when the unidentified object appears, it is one that is satisfied after the agent obtains the knowledge about whether to approach or avoid the stimulus. The value of this information is based not on the average of the obtainable outcomes, but is instead based on the knowledge of the effective outcomes – that the agent can either consume the positive reward or avoid the negative reward (see Fig. 2).

Finally, it is important to note that the opportunities to take particular actions (e.g., to orient) may themselves take on rewarding properties through some generalization or learning mechanism not included in this simulation. For example, the very act of orienting and determining "what's out there" could become rewarding to an organism based on the association between that action and the above-demonstrated emergent, always-positive reward-prediction error when new stimuli appear. A similar idea has been recently advanced by Redgrave and Gurney (2006) who hypothesize that an important purpose of the phasic dopamine response is to reinforce actions that occur before unpredicted salient events. The results here are not incompatible with that hypothesis, however it should be noted that Redgrave and Gurney's hypothesis is not directly tested in this simulation because no actions (i.e., exploration) were required of the agent in order for the salient event (the appearance of the object) to occur. However, the simulated phasic signal coincided with the time of the orienting response, suggesting that the two may be strongly related.

In closing, this article has demonstrated that RL principles can be used to explain a type of seemingly non-reward related activity of the dopaminergic neurons. This result emerged from the fact that the temporal-difference learning rule (such as that used by Kakade and Dayan (2002)) was embedded in a simulation in which the agent could select actions that had an effect on the eventual outcome. In the simulation, the agent learned that the outcome of orienting to an object that suddenly appeared could always either be rewarding or neutral because the negative outcome could be avoided. Therefore when the agent had an opportunity to orient, its reward-prediction error was always positive, computationally analogous to the novelty and saliency responses observed in biological organisms.

Acknowledgments

The work described in this article was supported by NIH R01 HD053639 and by NSF Training Grant DGE-9987588. I would like to thank Erik Reichle, Tessa Warren and an anonymous reviewer for helpful comments on an earlier version of this article.

References

Baird, L.C. (1995). Residual algorithms: Reinforcement Learning with function approximation. In A. Priedetis, & S. Russell (Eds.), *Machine learning: Proceedings of the twelfth international conference*.

Bunzeck, N., & Düzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron*, *51*(3), 369–379.

Dommett, E., Coizet, V., Blaha, C. D., Martindale, J., Lefebvre, V., Walton, N., et al. (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science*, *307*(5714), 1476–1479.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4–6), 495–506.

Gillies, A., & Arbuthnott, G. (2000). Computational models of the basal ganglia. *Movement Disorders*, *15*(5), 762–770.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, *96*(4), 651–656.

[4] One potential objection to the present work is that the orienting response appears to be hard-wired in the mammalian brain, for example, in projections from the superior colliculus (Dommett et al., 2005; Redgrave et al., 1999). In the present simulation, the agents were not hard-wired to orient to objects but instead learned an orienting behavior that permitted the eventual selection of an action (e.g., either approach or avoidance) that maximized reward. Similarly to hard-wired responses, these orienting behaviors occurred very rapidly, before the objects were identified, and were directed towards all objects. The goal of this work was not to make the claim that all such responses are learned, but rather that they can co-exist within the RL framework. Nevertheless, it would be interesting to investigate whether reward-related mechanisms might be involved in setting up connectivity in brainstem areas in order to generate this phasic dopamine response.

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, *15*(4–6), 549–559.
Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150–157.
McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346.
Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral and Brain Functions*, *1*, 6.
O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.
Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews Neuroscience*, *7*(12), 967–975.
Redgrave, P., Prescott, T. J., & Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends in Neurosciences*, *22*(4), 146–151.
Reichle, E. D., & Laurent, P. A. (2006). Using Reinforcement Learning to understand the emergence of "intelligent" eye-movement behavior during reading. *Psychological Review*, *113*(2), 390–408.
Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An introduction*. Cambridge, MA: MIT Press.
Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, *7*(8), 887–893.

Value-driven attentional capture

Brian A. Anderson[1], Patryk A. Laurent, and Steven Yantis

Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD 21218

Edited by Edward E. Smith, Columbia University, New York, NY, and approved May 16, 2011 (received for review March 11, 2011)

Attention selects which aspects of sensory input are brought to awareness. To promote survival and well-being, attention prioritizes stimuli both voluntarily, according to context-specific goals (e.g., searching for car keys), and involuntarily, through attentional capture driven by physical salience (e.g., looking toward a sudden noise). Valuable stimuli strongly modulate voluntary attention allocation, but there is little evidence that high-value but contextually irrelevant stimuli capture attention as a consequence of reward learning. Here we show that visual search for a salient target is slowed by the presence of an inconspicuous, task-irrelevant item that was previously associated with monetary reward during a brief training session. Thus, arbitrary and otherwise neutral stimuli imbued with value via associative learning capture attention powerfully and persistently during extinction, independently of goals and salience. Vulnerability to such value-driven attentional capture covaries across individuals with working memory capacity and trait impulsivity. This unique form of attentional capture may provide a useful model for investigating failures of cognitive control in clinical syndromes in which value assigned to stimuli conflicts with behavioral goals (e.g., addiction, obesity).

Effective deployment of attention is critical to the successful performance of any cognitive task. Attention determines what aspects of the sensory input are selected for cognitive processing, memory storage, and awareness. Two modes of attentional control are widely believed to determine perceptual priority: a voluntary, goal-directed mode, in which attention is guided by contextually appropriate goals and intentions, and an involuntary, stimulus-driven mode, in which attention is captured by physically salient stimuli (1–4) or by task-irrelevant stimuli that share identifying features with a searched-for target (5, 6). Each of these modes of control present concomitant benefits and costs: voluntary control of attention is goal-specific but potentially slower to implement; involuntary attentional capture can rapidly orient the organism to unexpected changes that could signal danger or opportunity, but has the potential to cause distraction from intended acts of perception.

Goal-directed and stimulus-driven modes of attentional control have long been a focus of intense investigation, and much has been learned about the operating principles of each mode of control and their interaction (1, 4). However, there is growing evidence that these are not the only influences on attentional deployment. To promote survival and well-being, the brain is optimized to learn about perceptual stimuli that signal the potential for procuring reward (7, 8). Voluntary attention to stimuli that predict reward is an effective mechanism for efficiently selecting valuable stimuli (9). Many studies have shown that reward facilitates voluntary attention to task-relevant stimuli, and that reward-based strategies and priorities strongly influence attentional performance (10–19).

Attentional capture by valuable but task-irrelevant stimuli could also confer adaptive advantages in many circumstances, leading the perceiver to orient to inconspicuous and/or unexpected reward-related stimuli. At the same time, however, attentional capture by reward-related stimuli (e.g., drugs of abuse, excessive food, or even irrelevant but rewarding information like an e-mail chime) can be maladaptive when it conflicts with contextually appropriate goals (e.g., intended abstinence from a drug or food) (20–25). This raises the possibility that valuable but inconspicuous stimuli capture attention involuntarily as a consequence of reward learning. Several recent studies have investigated this possibility, but in each case, when arbitrary stimuli were associated with reward delivery during a learning procedure and then made entirely task-irrelevant during extinction, they did not cause distraction (10–12, 26). Although it is known that task-irrelevant drug-related stimuli draw attention in addicted populations (27–30), and that motivationally salient stimuli, such as happy faces and erotic pictures, can capture attention (31, 32), it is unclear to what extent this reflects a general-purpose mechanism of attentional capture by stimuli imbued with value through reward learning. To date, no clear demonstration of such a mechanism of attentional capture, particularly in healthy individuals, has been reported.

We examined whether an irrelevant, unrewarded, and nonsalient distractor, previously associated with reward, captures attention when both stimulus-driven and goal-driven accounts predict that a physically salient and task-relevant target should instead solely determine the locus of attention. Critically, we associated value with a basic stimulus feature—color—rather than more complex conjunctions of features that have failed to capture attention during extinction in previous studies (10–12, 26). The results show that reward-related stimuli do cause significant and persistent distraction as a consequence of reward learning, and thereby reveal an involuntary mechanism of attentional selection that is uniquely value-driven, operating at an earlier level of representation than previously documented.

Results

Experiment 1. During an initial training phase, participants searched for a red or green target among differently colored nontargets (Fig. 1*A*), and received visual feedback at the end of each trial indicating an accumulating monetary reward for a correct response. Importantly, the participant's response did not depend on color; rather, they discriminated the orientation of a bar within the target stimulus; thus reward was associated with color, and not with a particular behavioral response. One target color (red for half the participants, green for the rest, to control for possible differences in physical salience) was associated with a high probability ($P = 0.8$) of a high reward (5¢) and a low probability ($P = 0.2$) of a low reward (1¢); for the other target color, this mapping was reversed. Participants were not explicitly informed of this reward contingency, but had to learn it over the course of 1,008 trials. Training thus imbued one color with high value and the other color with lower (but positive) value. After the training phase was complete, a test phase began, comprising 480 trials during which no reward was provided: participants searched for a unique shape in an array of six differently colored shapes (Fig. 1*B*). On half of these trials, one of the nontarget elements—termed the distractor—was rendered in red or green (each equally often); the target was never red or green, and participants

Author contributions: B.A.A., P.A.L., and S.Y. designed research; B.A.A. and P.A.L. performed research; B.A.A. and P.A.L. analyzed data; and B.A.A., P.A.L., and S.Y. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: bander33@jhu.edu.

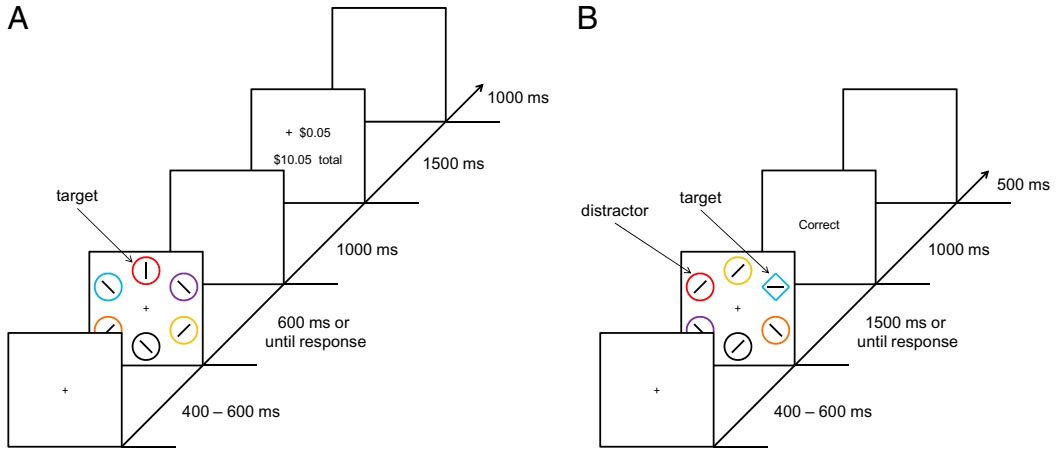

Fig. 1. Sequence of trial events. (*A*) Targets during the training phase were defined by color (red or green, exactly one of which was present on each trial), and participants reported the identity of the line segment inside of the target (vertical or horizontal). Correct responses were followed by the delivery of monetary reward feedback. One of the target colors was followed by a high reward on 80% of the trials and a low reward on 20% of the trials; for the other target color, this mapping was reversed. (*B*) During the test phase, the target was defined as the unique shape, and no reward feedback was provided. On half of the trials, one of the nontarget items—the distractor—was rendered in the color of a formerly rewarded target (each color equally often).

were informed that color was irrelevant to the task and should be ignored.

Based on the reward contingencies to which participants were exposed during the training phase, trials during the test phase were classified as containing a high-value distractor, a low-value distractor, or neither. A repeated-measures ANOVA revealed that response times (RTs) differed significantly among these three conditions [$F(2, 50) = 6.07$, $P = 0.004$] (Table 1). High-value distractors slowed RT relative to when neither value-related distractor was present [$t(25) = 3.49$, $P = 0.002$], and the effect of reward on performance exhibited a linear trend [$F(1, 25) = 12.19$, $P = 0.002$]. There was no significant difference in error rate between the three distractor conditions [$F(2, 50) = 0.41$, $P = 0.667$]. These results are striking in that they clearly violate the predictions of both a salience-driven and goal-driven account of attentional capture: the data mirror the well-documented distracting effect of physically salient stimuli (2, 4, 33), despite the fact that the distractors were neither physically salient nor goal-relevant, and did not have any identifying features in common with the searched-for target (5). Even the fastest 25% of RTs in the high-value distractor condition were slower than those in the distractor-absent condition [$t(25) = 3.07$, $P = 0.005$], suggesting that the high-value distractor captured attention consistently, rather than on only a small proportion of the trials (34). To confirm that red and green were not more physically salient than the other colors we used (and equally salient to each other), eight naive participants completed a separate control experiment that consisted only of the test phase of experiment 1. Mean RT did not differ for the three conditions containing a red distractor, a green distractor, or neither [$F(2, 14) = 0.13$, $P = 0.880$] (Table 2).

Table 1. Mean response time (in milliseconds) and error rate, respectively, in the test phase of the experiments in which reward was delivered for each of three training conditions: long training (1,008 trials) with low and high reward of 1¢ or 5¢ per trial, brief training (240 trials) with rewards of 2¢ and 10¢ per trial, and brief training followed by a delay of 4–21 d

	Distractor condition in the test phase		
Training phase	None	Low value	High value
	665 (2.8)	673 (2.8)	681 (2.6)
1,008 trials	0.11 (0.004)	0.10 (0.004)	0.11 (0.004)
	667 (2.0)	675 (3.0)	682 (2.9)
240 trials	0.12 (0.005)	0.12 (0.006)	0.12 (0.006)
	614 (1.8)	624 (2.7)	630 (3.3)
4–21 d ago	0.06 (0.004)	0.07 (0.006)	0.08 (0.005)

The error terms, in parentheses, reflect the within-subjects SEM.

Visual working memory (WM) capacity is correlated with the magnitude of attentional capture by salient, task-irrelevant stimuli (35, 36), an effect that is thought to reflect individual differences in a general ability to resist distraction and thus to maintain items in working memory. Trait impulsivity also provides a broad measure of the degree to which people can successfully exercise control over behavior (37). We therefore measured visual working memory capacity using a change-detection task (35) and trait impulsivity using the Barratt Impulsiveness Scale (BIS-11) (38), and entered them as predictors of value-driven attentional capture in a simultaneous regression model. These two factors accounted well for individual differences in value-driven attentional capture ($R^2 = 0.355$, $P = 0.006$); each variable uniquely contributed to the predictive power of the model ($\beta = -0.554$, $P = 0.004$, and $\beta = 0.378$, $P = 0.038$, for WM capacity and impulsivity, respectively; Fig. 2). Greater impulsivity and smaller working memory capacity were associated with greater vulnerability to distraction by high-value stimuli. The relationship between visual working memory capacity and the magnitude of value-driven capture cannot be explained by a general slowing of response times by lower-capacity individuals, because visual working memory capacity was not correlated with RT in the no-distractor condition (Pearson's $r = 0.073$, $P = 0.723$).

Experiment 2. An alternative account for our findings is that participants deliberately continued to search for the red and green items in the test phase even though those items were no longer task-relevant or rewarded. Although it is known that attentional priorities are rapidly adjusted with changing task demands (39), former targets can attract attention under certain circumstances (40, 41). To rule out this account, we tested 10 new participants who engaged in the same training and test phases of the experiment, but with no reward feedback during training. Instead, participants were compensated with a flat rate that matched the average earnings of participants in the main experiment ($25). We found that removing trial-by-trial reward feedback from the training phase completely abolished any effect of distractors at test [$t(9) = -0.39$, $P = 0.707$] (Table 2). There was

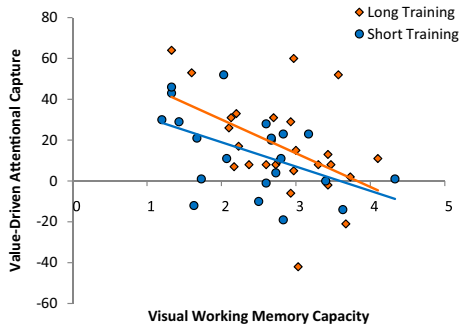

Fig. 2. Magnitude of value-driven attentional capture (indexed as response time when a high-value distractor was present minus response time when no distractor was present) as a function of visual working memory capacity following the long (yellow) and short (blue) training regimen. Pearson product-moment correlations for the long and short training experiments were −0.468 ($P = 0.016$) and −0.468 ($P = 0.021$), respectively.

no significant difference in error rate between the red, green, and no-distractor conditions [$F(2, 18) = 2.30$, $P = 0.139$].

Experiment 3. To assess the robustness of value-driven attentional capture, we shortened the training phase and test phase to 240 trials each and doubled the reward magnitude (2¢ and 10¢, respectively, for low and high reward) with 24 new participants. We replicated the effect of prior reward on performance during the test phase [repeated-measures ANOVA: $F(2, 46) = 5.17$, $P = 0.009$; linear trend: $F(1, 23) = 14.09$, $P = 0.001$] (Table 1). The magnitude of slowing caused by high-value distractors was again correlated with visual working memory capacity (Pearson's $r = -0.468$, $P = 0.021$; Fig. 2), but not with trait impulsivity (Pearson's $r = 0.093$, $P = 0.666$). Again, there was no correlation between visual working memory capacity and RT in the no-distractor condition (Pearson's $r = 0.027$, $P = 0.900$). We then invited these participants back after several days had passed (4–21 d, mean = 8.8, SD = 3.9) to carry out the same test phase with no additional training; 19 of the 24 participants returned. The influence of prior reward on performance remained [repeated-measures ANOVA: $F(2, 36) = 5.81$, $P = 0.007$; linear trend: $F(1, 18) = 11.55$, $P = 0.003$]. In none of these experiments did error rates differ significantly among the three conditions (all Ps > 0.25). These results show that high-value stimuli have a rapidly learned and long-lasting influence on attentional priority, and the magnitude of this effect is negatively correlated with working memory capacity. The influence of trait impulsivity was only evident for well-learned associations provided by the longer training regimen of the first experiment, which may be a manifestation of habit learning (42).

Finally, we tested whether value-driven attentional capture entails a spatial deployment of attention to the distractor's location or, instead, a nonspecific filtering cost—that is, an overall slowing caused by the presence of a high-value stimulus. Previous studies have shown that attentional capture by a physically salient distractor inhibits perception of a stimulus that subsequently appears in the distractor's location (33). We examined response times in trials on which no distractor was presented that were preceded by a trial containing a high-value distractor. Responses were on average 66 ms slower when the target appeared in a location formerly occupied by a high-value distractor than when it appeared in another location [$t(23) = 3.13$, $P = 0.005$], confirming that high-value distractors indeed capture attention in a spatially specific manner.

Table 2. Mean response time (in milliseconds) and error rate, respectively, in the test phase of the experiments in which no reward was delivered

	Distractor condition in the test phase		
Training phase	None	Red	Green
None	698 (4.1) 0.13 (0.004)	696 (4.7) 0.13 (0.006)	700 (3.4) 0.14 (0.006)
1,008 trials (unrewarded)	602 (3.9) 0.14 (0.004)	606 (2.1) 0.17 (0.006)	593 (3.9) 0.15 (0.005)

The error terms, in parentheses, reflect the within-subjects SEM.

Discussion

Two modes of attentional control have long been known to play a role in the adaptive deployment of selective attention (1–6). Several recent studies have shown that the voluntary deployment of attention is influenced by reward (10–19). In contrast to the ample evidence that voluntary deployment of attention to task-relevant stimuli is affected by reward, the evidence for an influence of the value assigned to stimuli through reward learning on involuntary attentional capture is negative or equivocal (10–12, 26). The experiments reported here provide clear evidence that arbitrary reward-related stimuli capture attention involuntarily and persistently as a result of associations that develop rapidly during learning.

Value-driven attentional capture is distinct from the well-established role of salience and ongoing goals in the control of attention. Our findings establish that nonsalient, task-irrelevant stimuli previously associated with reward slow visual search during extinction, and that the magnitude of slowing is spatially specific: when a target appears in a location occupied by a high-value distractor on the previous trial, slowing is especially prolonged. This result strongly suggests that high-value distractors draw spatial attention, and the subsequent act of disengagement leaves an inhibitory trace at that location (33). Furthermore, a control experiment showed that the effect could only be attributable to reward feedback during training, ruling out the persistence of a deliberate attentional strategy as an explanation; this confirms a recent report that top-down goals in visual search can be adjusted flexibly within seconds (39), and further distinguishes value-driven capture from goal-directed attentional deployment.

In previous studies that have investigated whether valuable stimuli capture attention as a consequence of reward learning (10–12, 26), the types of stimuli used were complex (e.g., faces and words). We observed evidence of value-driven attentional capture operating on a basic stimulus feature—color—that can provide a basis for the efficient detection of valuable stimuli. This is consistent with the level of selectivity seen in contingent involuntary orienting (5, 6), and may reflect a general underlying principle of involuntary attentional selection.

In a classic investigation of attentional control, Shiffrin and Schneider (41) reported that following extensive training in search for a specific target letter, that letter subsequently captures attention even when it is no longer task-relevant. However, this effect required a great deal of consistent training. The significant slowing caused by value-driven attentional capture reported here required as little as 240 trials during training. Furthermore, if reward was omitted during training, the effect disappeared. Value-driven capture is clearly a distinct phenomenon.

Individual differences in visual working memory capacity are thought to reflect variation in a general ability to resist distraction; in a change-detection task, this manifests as the efficient and selective processing of a capacity-limited number of to-be-remembered items, with minimal interference from irrelevant and supracapacity items (35, 36). Individual differences in change-detection performance thus reflect variation in the ability to restrict visual and mnemonic processing to goal-relevant features and locations. We found that individual differences in visual working memory capacity are strongly correlated with suscepti-

bility to value-driven attentional capture: individuals with low working memory capacity tend to exhibit stronger value-driven attentional capture. This finding echoes recent reports that WM capacity is correlated with the degree to which physically salient, contingently relevant stimuli capture attention despite competing goals (35, 36), and further extends previous demonstrations of a correlation between verbal measures of working memory capacity and the efficiency of goal-directed attentional selection (43–45).

The influence of reward in motivated behavior has been a focus of intense investigation in recent years. A wide variety of stimuli are rewarding—sweet taste, positive facial expression, erotic pictures, money, and illicit drugs such as cocaine, among others. The receipt of a rewarding stimulus is accompanied by subjective pleasure, and associative learning mechanisms in the brain give rise to incentive salience—a desire or "wanting" response when reward-associated stimuli are present (22). In susceptible individuals, the learned wanting response can override cognitive intentions to avoid the rewarding stimulus, and lead to impairment of cognitive control and ultimately to addiction and related syndromes.

Although it is known that irrelevant drug-related stimuli draw attention in addicted populations (27–30), it is unclear to what degree such effects might be explained by attentional capture driven by associations between stimuli and reward that arise through associative learning. Drugs of abuse usurp the brain's reward system, making drug addiction more than just a consequence of normal reward learning. Motivationally salient stimuli, such as happy faces and erotic pictures, are also known to capture attention (31, 32), but it is unclear whether such attentional preferences reflect arbitrary associations between stimuli and reward outcomes that develop through basic learning processes, or an evolutionarily conserved attentional bias. Our results clearly demonstrate that learned stimulus-reward associations are sufficient to involuntarily drive attention allocation, suggesting that maladaptive attentional biases found in drug addiction (27–30) may reflect, in part, the disordered influence of an otherwise normal cognitive process.

Value-driven attentional capture may play a key role in a variety of clinical syndromes in which both attention and reward have been critically implicated, including drug addiction (20–22), obesity (23), attention-deficit hyperactivity disorder (24), and obsessive-compulsive disorder (25). These conditions are highly comorbid (23, 25), suggesting common underlying causal factors. We observed clear individual differences in, and patterns of correlation with, the magnitude of value-driven attentional capture: individuals with low visual working memory capacity and high trait impulsivity were the most vulnerable to the effect of stimulus value on involuntary attentional selection. These individual differences may provide insights into the traits and states that jointly influence susceptibility to these conditions.

Methods

Experiment 1. ***Participants.*** Twenty-six participants were recruited from the Johns Hopkins community. All were screened for normal or corrected-to-normal visual acuity and color vision. Participants were provided monetary compensation based on performance (mean = $25.11), in addition to $5 compensation for completing an initial session. Informed consent was obtained from all participants, and all of the experimental procedures were approved by the Johns Hopkins University Institutional Review Board.

Initial session. Participants first filled out a questionnaire (BIS-11) and performed a change-detection task designed to measure visual working memory capacity. The questionnaire and change-detection task were performed the day before the experiment in a single 20-min session. The method for the change-detection task has been previously described (35). In the change-detection task, participants were presented with a brief unmasked display consisting of four, six, or eight colored squares, which was followed by a probe display in which a single colored square appeared in a previously occupied location. Participants indicated, via an unspeeded key press, whether this probe square was the same or different in color than the square previously presented in the probed location. Visual working memory capacity was measured separately for each display size by multiplying the display size by the difference between the hit rate and false alarm rate, and then averaged across set sizes to obtain a global estimate (35).

Apparatus. A Mac Mini equipped with Matlab software and PsychToolbox extensions was used to present the stimuli on a Dell P991 monitor. The participants viewed the monitor from a distance of ~50 cm in a dimly lit room. Reponses were entered by using a standard 101-key US layout keyboard.

Stimuli. The sequence of events and time course for the training and test phases are shown in Fig. 1 *A* and *B*, respectively. Each trial consisted of three displays: a fixation display, a search display, and a feedback display. During both the training and test phases, the fixation display consisted of a white fixation cross (0.5° × 0.5° visual angle) presented in the center of the screen against a black background, and the search display consisted of the fixation cross surrounded by six shapes (2.3° × 2.3° visual angle) placed at equal intervals along an imaginary circle with a 5° radius.

Training phase. During the training phase, the six shapes in the search display were all circles of different colors (red, green, blue, cyan, pink, orange, yellow, and white). Targets were defined as a red or a green circle, exactly one of which was presented on every trial. Inside the target, a white line segment was oriented either vertically or horizontally, and inside each of the nontargets, a white line segment was tilted at 45° to the left or to the right. The feedback display informed participants of the reward earned on the previous trial, as well as total reward accumulated thus far.

Test phase. During the test phase, the search display consisted of a circle among diamonds or a diamond among circles, and the target on each trial was defined as the unique shape. Each item in the display had a unique color. On half of the trials, one of the nontarget elements, the distractor, was rendered in red or green; the target was never red or green, and participants were informed that color was irrelevant to the task and should be ignored. The feedback display at test informed participants only whether their response on the previous trial was correct. That is, no reward was provided during the test phase.

Design. The experiment consisted of 1,008 training trials and 480 test trials. Participants were provided with 50 practice trials before the training trials, and 20 practice (distractor absent) trials before the test trials. The practice trials were identical to the experimental trials except that no reward feedback was provided to the participants. After each 100 experimental trials and between the tasks, participants were provided with a short break. Target identity, target location, distractor color, and distractor location were fully crossed and counterbalanced.

Correct responses were followed by visual feedback indicating monetary reward in the training phase. High-reward targets were followed by high-reward feedback (5¢) on 80% of trials and low-reward feedback (1¢) on the remaining 20%; for low-reward targets, the percentages were reversed. High-reward targets were red for half of the participants, and green for the other half. No reward feedback was provided during the initial practice block, and no reward feedback was provided during the test phase. Upon completion of the experiment, participants were given the cumulative monetary reward they had earned.

Procedure. Each participant was tested individually over the course of a single 2-h session. Each session took place inside a dimly lit laboratory room. The experimenter familiarized all participants with each task by providing written and oral descriptions of the stimuli and procedures. Participants were instructed to respond "as quickly as possible while minimizing errors."

Each trial began with the presentation of the fixation display for a randomly varying interval of 400, 500, or 600 ms. The search display then appeared and remained on screen until a response was made or the trial timed out. The training task was performed under time pressure, with trials terminating after 600 ms; during the test, time pressure was lifted by lengthening this time limit to 1,500 ms.

Participants made a forced-choice target identification by pressing the *z* and *m* keys for the vertically and horizontally orientated targets, respectively. Response time was measured from the onset of the target display until a response was made or the trial timed out. The computer emitted a 500-ms, 1,000-Hz feedback tone to inform the participant when a trial timed out. Only correct responses were included in the analysis, and all RTs more than three SDs above and below the mean for each subject and condition were excluded from the analysis.

Experiment 2. ***Participants.*** Ten participants were recruited from the Johns Hopkins community. All were screened for normal or corrected-to-normal visual acuity and color vision. Participants were compensated with a flat amount of $25. None of the participations had participated in experiment 1.

Apparatus and stimuli. The apparatus and stimuli were identical to experiment 1, with the exception that the feedback display during training informed

participants only whether their previous response was correct. Critically, no reward feedback was provided.

Design and procedure. The design and procedure were identical to experiment 1, with the exception that no monetary rewards were provided for correct responses. Also, participants did not perform the change detection task or complete the BIS-11.

Experiment 3. ***Participants.*** Twenty-four new participants were recruited from the Johns Hopkins community. All were screened for normal or corrected-to-normal visual acuity and color vision. Participants were provided monetary compensation based on performance (mean = $13.24).

Apparatus and stimuli. The apparatus and stimuli were identical to experiment 1.

Design and procedure. The design and procedure were identical to experiment 1, with the following exceptions. The experiment consisted of a single 1-h session. The training and test phases consisted of 240 trials each, with a short break every 120 trials. Trials terminated after 800 ms in the training phase and 1,200 ms in the test phase. High and low rewards were increased to 10¢ and 2¢, respectively. Participants performed the change detection task and completed the BIS-11 at the beginning of the session. Four to 21 d after the initial session (mean = 8.8, SD = 3.9), 19 of the participants returned to complete the test phase again, and were compensated with an additional $5.

ACKNOWLEDGMENTS. We thank T. Braver, H. Egeth, J. Flombaum, L. Gmeindl, S. Grant, P. Holland, D. E. Meyer, J. Serences, and D. Strayer for comments and suggestions, and E. Wampler for help in data collection. This research was supported by National Institutes of Health Grant R01-DA013165 (to S.Y.).

1. Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* 3:201–215.
2. Theeuwes J (1992) Perceptual selectivity for color and form. *Percept Psychophys* 51: 599–606.
3. Itti L, Koch C (2001) Computational modelling of visual attention. *Nat Rev Neurosci* 2: 194–203.
4. Theeuwes J (2010) Top-down and bottom-up control of visual selection. *Acta Psychol (Amst)* 135:77–99.
5. Folk CL, Remington RW, Johnston JC (1992) Involuntary covert orienting is contingent on attentional control settings. *J Exp Psychol Hum Percept Perform* 18:1030–1044.
6. Anderson BA, Folk CL (2010) Variations in the magnitude of attentional capture: Testing a two-process model. *Atten Percept Psychophys* 72:342–352.
7. Shuler MG, Bear MF (2006) Reward timing in the primary visual cortex. *Science* 311: 1606–1609.
8. Seitz AR, Kim D, Watanabe T (2009) Rewards evoke learning of unconsciously processed visual stimuli in adult humans. *Neuron* 61:700–707.
9. Maunsell JHR (2004) Neural representations of cognitive state: Reward or attention? *Trends Cogn Sci* 8:261–265.
10. Della Libera C, Chelazzi L (2009) Learning to attend and to ignore is a matter of gains and losses. *Psychol Sci* 20:778–784.
11. Raymond JE, O'Brien JL (2009) Selective visual attention and motivation: The consequences of value learning in an attentional blink task. *Psychol Sci* 20:981–988.
12. Krebs RM, Boehler CN, Woldorff MG (2010) The influence of reward associations on conflict processing in the Stroop task. *Cognition* 117:341–347.
13. Della Libera C, Chelazzi L (2006) Visual selective attention and the effects of monetary reward. *Psychol Sci* 17:222–227.
14. Hickey C, Chelazzi L, Theeuwes J (2010) Reward changes salience in human vision via the anterior cingulate. *J Neurosci* 30:11096–11103.
15. Hickey C, Chelazzi L, Theeuwes J (2010) Reward guides vision when it's your thing: Trait reward-seeking in reward-mediated visual priming. *PLoS ONE* 5:e14087.
16. Peck CJ, Jangraw DC, Suzuki M, Efem R, Gottlieb J (2009) Reward modulates attention independently of action value in posterior parietal cortex. *J Neurosci* 29:11182–11191.
17. Serences JT (2008) Value-based modulations in human visual cortex. *Neuron* 60: 1169–1181.
18. Navalpakkam V, Koch C, Rangel A, Perona P (2010) Optimal reward harvesting in complex perceptual environments. *Proc Natl Acad Sci USA* 107:5232–5237.
19. Pessoa L, Engelmann JB (2010) Embedding reward signals into perception and cognition. *Front Neurosci*, 10.3389/fnins.2010.00017.
20. Garavan H, Hester R (2007) The role of cognitive control in cocaine dependence. *Neuropsychol Rev* 17:337–345.
21. Field M, Cox WM (2008) Attentional bias in addictive behaviors: A review of its development, causes, and consequences. *Drug Alcohol Depend* 97:1–20.
22. Robinson TE, Berridge KC (2008) Review. The incentive sensitization theory of addiction: some current issues. *Philos Trans R Soc Lond B Biol Sci* 363:3137–3146.
23. Davis C (2010) Attention-deficit/hyperactivity disorder: Associations with overeating and obesity. *Curr Psychiatry Rep* 12:389–395.
24. Bush G (2010) Attention-deficit/hyperactivity disorder and attention networks. *Neuropsychopharmacology* 35:278–300.
25. Sheppard B, et al. (2010) ADHD prevalence and association with hoarding behaviors in childhood-onset OCD. *Depress Anxiety* 27:667–674.
26. Rutherford HJV, O'Brien JL, Raymond JE (2010) Value associations of irrelevant stimuli modify rapid visual orienting. *Psychon Bull Rev* 17:536–542.
27. Stormark KM, Field NP, Hugdahl K, Horowitz M (1997) Selective processing of visual alcohol cues in abstinent alcoholics: An approach-avoidance conflict? *Addict Behav* 22:509–519.
28. Lubman DI, Peters LA, Mogg K, Bradley BP, Deakin JFW (2000) Attentional bias for drug cues in opiate dependence. *Psychol Med* 30:169–175.
29. Field M, Mogg K, Zetteler J, Bradley BP (2004) Attentional biases for alcohol cues in heavy and light social drinkers: The roles of initial orienting and maintained attention. *Psychopharmacology (Berl)* 176:88–93.
30. Field M, Mogg K, Bradley BP (2004) Eye movements to smoking-related cues: Effects of nicotine deprivation. *Psychopharmacology (Berl)* 173:116–123.
31. Hodsoll S, Viding E, Lavie N (2011) Attentional capture by irrelevant emotional distractor faces. *Emotion* 11:346–353.
32. Most SB, Smith SD, Cooter AB, Levy BN, Zald DH (2007) The naked truth: Positive, arousing distractors impair rapid target perception. *Cogn Emotion* 21:964–981.
33. Theeuwes J, Godljn R (2002) Irrelevant singletons capture attention: Evidence from inhibition of return. *Percept Psychophys* 64:764–770.
34. Yantis S, Meyer DE, Smith JEK (1991) Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychol Bull* 110:350–374.
35. Fukuda K, Vogel EK (2009) Human variation in overriding attentional capture. *J Neurosci* 29:8726–8733.
36. Fukuda K, Vogel EK (2011) Individual differences in recovery time from attentional capture. *Psychol Sci* 22:361–368.
37. Dickman SJ, Meyer DE (1988) Impulsivity and speed-accuracy tradeoffs in information processing. *J Pers Soc Psychol* 54:274–290.
38. Patton JH, Stanford MS, Barratt ES (1995) Factor structure of the Barratt impulsiveness scale. *J Clin Psychol* 51:768–774.
39. Lien M-C, Ruthruff E, Johnston JC (2010) Attentional capture with rapidly changing attentional control settings. *J Exp Psychol Hum Percept Perform* 36:1–16.
40. Kyllingsbaek S, Schneider WX, Bundesen C (2001) Automatic attraction of attention to former targets in visual displays of letters. *Percept Psychophys* 63:85–98.
41. Shiffrin RM, Schneider W (1977) Controlled and automatic human information processing II: Perceptual learning, automatic attending, and general theory. *Psychol Rev* 84:127–190.
42. Wood W, Neal DT (2007) A new look at habits and the habit-goal interface. *Psychol Rev* 114:843–863.
43. Bleckley MK, Durso FT, Crutchfield JM, Engle RW, Khanna MM (2003) Individual differences in working memory capacity predict visual attention allocation. *Psychon Bull Rev* 10:884–889.
44. Kane MJ, Bleckley MK, Conway ARA, Engle RW (2001) A controlled-attention view of working-memory capacity. *J Exp Psychol Gen* 130:169–183.
45. Kane MJ, Engle RW (2003) Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *J Exp Psychol Gen* 132:47–70.

A Neural Mechanism for Reward Discounting: Insights from Modeling Hippocampal-Striatal Interactions

Patryk A. Laurent

Corresponding author's address:

Department of Psychological and Brain Sciences

The Johns Hopkins University, 136 Ames Hall

3400 N. Charles St, Baltimore MD, 21218

Email: laurent@jhu.edu

Telephone: +1-410-929-2562

Fax: +1-410-516-4478

Abstract

Decision-making often requires taking into consideration immediate gains as well as delayed rewards. Studies of behavior have established that anticipated rewards are discounted according to a decreasing hyperbolic function. Although mathematical explanations for reward delay discounting have been offered, little has been proposed in terms of neural network mechanisms underlying discounting. There has been much recent interest in the potential role of the hippocampus. Here we demonstrate that a previously-established neural network model of hippocampal region CA3 contains a mechanism that could explain discounting in downstream reward-prediction systems (e.g., basal ganglia). As part of its normal function, the model forms codes for stimuli that are similar to future, predicted stimuli. This similarity provides a means for reward predictions associated with future stimuli to influence current decision-making. Simulations show that this "predictive similarity" decreases as the stimuli are separated in time, at a rate that is consistent with hyperbolic discounting.

Keywords: hippocampus, reward discounting, neural network, prediction

Introduction

A hallmark of intelligent decision-making is the ability to balance short term gains against long-term rewards. For example, is $400 now better than $1000 in a year? It seems like a better deal to wait a year for the $1000. However, the answer to the question may become less clear if we are comparing $400 now to $1000 in five years. Studies have now shown that decisions about reward made by pigeons, rats, as well as humans, can be described by hyperbolic discounting functions (Kacelnik, 1997; Mazur, 1987; Mazur & Biondi, 2009). Measuring such functions empirically involves determining the point at which organisms chose between two options, e.g., an immediate small reward and a delayed large rewards, with equal probability. A typical study with pigeons involves observing as they learn to press either a red button or a blue button, each of which leads to a different fixed amount of grain (e.g., red = grain for 2 seconds versus blue = grain for 6 seconds). By adjusting the delay to the larger reward, experimenters can observe when the pigeons choose the red and blue buttons with equal probability, that is, the point at which the immediate small reward and delayed large reward are treated as equivalent.

Going back to our example for illustrative purposes, suppose we have an imaginary participant whose hyperbolic discounting function is $V(t) = \alpha/(1 + 0.29t)$, where α is the offered reward, t is the elapsed time (here, in years), and $V(t)$ is the discounted value. Then for our participant, $1000 in

year is worth $775.19 which is greater than the $400 now, and our participant would choose the $1000. However, using the same equation, $1000 in 5 years would be worth only $408.16 which is very close to the $400 now option – so the decision becomes much less clear. Indeed for participants with steeper discounting functions, as is this case with children (Green et al., 1999), the $400 now is selected significantly more often. Decision-making theories like Reinforcement Learning successfully incorporate this aspect of decision making using a parameter known as the "discount factor" (Sutton & Barto, 1998).

Although reward delay discounting is well established in the behavioral literature, proposals as to how neural networks in the brain give rise to the discounting function or to its hyperbolic shape are lacking. Researchers have, however, implicated particular brain regions in reward delay discounting. One idea is that discounting emerges from interactions within a single prefrontal-posterior parietal-basal ganglia system (Kable & Glimcher, 2007, 2010), or between ventromedial prefrontal cortex and lateral prefrontal cortex (Figner et al., 2010; Hare et al., 2009). Another idea is that discounting reflects a tradeoff between the prefrontal cortex and particular limbic regions. In those accounts, prefrontal regions are involved in procuring delayed rewards whereas limbic regions are involved in immediate rewards (McClure et al., 2007, 2004).

Of particular interest is a third idea, which is based on the finding that hippocampal lesions increase the preference for immediate rewards, i.e., in-

creased discounting (Cheung & Cardinal, 2005; Gupta et al., 2009; Mariano et al., 2009; Rawlins et al., 1985). Here, the suggestion is that discounting may be mediated by an interaction, rather than a trade-off, between prefrontal and hippocampal regions. A possible pathway for this interaction is documented in existing anatomy and neurophysiology research: The hippocampus projects to the ventral striatum (Kelley & Domesick, 1982), and *in vivo* work suggests strong functional connectivity between these regions (Lansink et al., 2009; Pennartz et al., 2004).

The case for hippocampal contributions to reward delay discounting is bolstered by data from a recent neuroimaging study. In participants with intact brains, "episodic future thinking" *reduced* reward discounting such that future rewards were valued more highly. (Episodic future thinking refers to an experimental manipulation which arguably directs attention to the hippocampal formation, which is thought to be involved in episodic cognitive representations; see Peters & Büchel, 2010). The decrease in reward discounting was accompanied by increased functional connectivity between the prefrontal cortex and the hippocampus. This suggests that an increase in hippocampal activity results in an overall increase in the value of future (delayed) reward predictions.

Despite all of these findings, the question remains: how does the hippocampus contribute to decision-making about anticipated rewards? The main possibility discussed in the existing literature is one called future "mental simulation". This idea is based on findings that the hippocampus ap-

pears to be required for the imagination of future events (Addis et al., 2010; Gamboz et al., 2010; Hassabis et al., 2007; Schacter & Addis, 2009). These "mental simulations" putatively allow an organism to evaluate the value of future events (Johnson et al., 2007; Johnson & Redish, 2007). Although it is straightforward to see how mental simulation could be used to decide between two options with eventual outcomes of differing value, it is less obvious whether mental simulation allows an organism to discount delayed rewards and thereby make decisions that take intervening time into consideration.

Here we propose an account of hippocampal involvement in reward delay discounting based on a previously formulated model of hippocampal region CA3 established by Levy and colleagues (Levy, 1989, 1996; Levy et al., 2005). That model states that region CA3 of the hippocampus re-encodes representations of stimuli or events to become more similar to the events that they predict. This leads to representations in the present that partially activate representations for those future events. We posit that these partial reactivations cause activation in downstream reward-predictive regions, namely in the striatum, resulting in the generation of discounted reward predictions. An analysis shows that this "predictive similarity" decreases in time in a manner that approximates hyperbolic discounting. Finally, we operationalize a mental simulation account of reward delay discounting, and attempt to test its predictions for compatibility with the model.

Methods

Reward-Prediction Component

The first component of the model is a two-layer network that generates "bottom-up" reward predictions for a sensory stimulus. In the network, when an activity pattern x (e.g., a visual stimulus) is presented on the input layer of this network, its output layer represents the amount of reward the organism expects to subsequently receive, $V(x)$. This component is inspired by the idea that reward predictions might be learned through the strengthening of projections from neocortex to the striatum by phasic dopamine activity in the presence of unexpected reward (Bamford et al., 2004; Reynolds & Wickens, 2002). This manner of modeling cortico-striatal projections is common in actor-critic reinforcement learning architectures (Joel et al., 2002; Suri & Schultz, 1998). After learning is completed, stimuli that have become associated with reward evoke reward prediction responses in the striatum (Hollerman et al., 1998; Schultz et al., 1997).

To generate its reward predictions, the network computes a weighted sum of each element of the x vector multiplied by its corresponding element in a weight vector w and then optionally applying a final function f which could be either linear or nonlinear. Thus the resulting output is the model's predicted reward for the stimulus, $V(x) = f(w \cdot x)$. Here we assume a linear output for simplicity.

Discounting Component

The second component of the model, comprising a recurrent network and a decoder network, is the source of the representations that lead to discounted reward predictions. This component is based on a minimalist biologically-inspired neural network model of the hippocampus (Levy, 1989, 1996; Levy et al., 2005). The model's architecture and usage are inspired by a range of anatomical, functional, and neurophysiological findings in hippocampus. The model and its accompanying theory have been previously used to develop mechanistic explanations for putatively hippocampally-dependent phenomena in the brain like orthogonalization of similar episodic experiences, the generation of sparse representations, and the temporal compression of replayed sequences during sleep (see Levy, 1996, for a review).

The hippocampal model is centered around a putative CA3 network consisting of a large number of randomly, sparsely connected, neuron-like units organized in an activity-controlled recurrent network. The neuron-like units are McCulloch-Pitts (binary or "on-off") units, but more realistic integrate-and-fire units can also be used (August & Levy, 1999). During each simulated time-step, the simulated neurons integrate inputs on their dendrites and then emit a spike if the weighted sum is sufficiently large according to a dynamic threshold. The excitation y of each unit j on time-step t is modeled as a weighted sum of the binary output of all of its afferent input units on the previous time-step,

$$y_j(t) = \sum_{i=1}^{n} w_{ij} \cdot Z_i(t-1). \tag{1}$$

Firing of neuron j ($Z_j = 1$) occurs if the excitation for that neuron exceeds a network-wide dynamic threshold θ_t, or if the neuron is directly stimulated (i.e., clamped) by an external input $x_j(t)$:

$$Z_j(t) = \begin{cases} 1 & \text{if } y_j(t) > \theta_t \text{ or } x_j(t) = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

This time-varying threshold qualitatively allows for a k-winners-take-all (or "competitive") activity control. Such activity control is intended to operationalize the existence in CA3 of an extensive network of inhibitory interneurons that keep the network-wide activity to an approximately constant level.

The strength of the synaptic connections between the neurons, or weights, w_{ij}, range from 0 and 1. Because the network is sparsely connected, the majority of those weights are 0 to indicate that neurons share no connections.

Learning in the network occurs according to a temporally-asymmetric Hebbian-like learning rule inspired by spike-timing dependent plasticity or STDP, which has been observed in the form of temporal contiguity requirements for long-term potentiation (Bi & Poo, 1998; Levy & Steward, 1983; Markram et al., 1997). Each weight from neuron i to neuron j is updated only on time-steps when the post-synaptic neuron j is active (i.e., $Z_j = 1$). The update depends on the difference between the current weight w_{ij} and a

running average of afferent presynaptic activity,

$$w_{ij}(t+1) = w_{ij}(t) + \mu \cdot Z_j(t)\Big(\bar{Z}_i(t-1) - w_{ij}(t)\Big), \tag{3}$$

where the running average saturates when the presynaptic neuron fires, and decreases exponentially otherwise, i.e.,

$$\bar{Z}_i(t) = \begin{cases} 1 & \text{if } Z_i(t) = 1, \\ \alpha \cdot \bar{Z}_i(t-1) & \text{otherwise.} \end{cases} \tag{4}$$

The constant μ is a network-wide synaptic weight modification rate constant (i.e., "learning rate"). The use of a running average of afferent presynaptic activity for weight modification is intended to mimic a saturate-and-decay model of the NMDA receptor. Equation 4 indicates that the running average of firing activity is updated on every time-step for each neuron i. The quantity α is a time-constant of the decay (i.e., of the modeled NMDA receptors). Although real time and simulated time can be equated through careful tuning of the α parameter as per Mitman et al. (2003), the present modeling work does not make use of time in the absolute sense, only in the relative sense. That is, no attempt is made to model real time.

Synaptic plasticity mechanisms between the units cause neurons that initially fire for a given stimulus to gradually, through learning, begin firing for stimuli that anticipate the later stimuli (Levy et al., 2005; Mitman et al., 2003). An analog of this "earlier-shifting" has been observed to occur in

recordings of rat hippocampus (e.g., Mehta et al., 1997, 2000), and is the basis for the "predictive similarity" account of reward discounting proposed in this article.

Each network consisted of 4,096 neurons with a random connection probability of 8% (self-connections were not allowed) and activity level was regulated to be 7.5% through a winners-take-all competitive mechanism for simplicity, as controlled by θ from Equation 2. Each network was trained for 120 trials on a sequence of 20 orthogonal (non-overlapping) stimuli each activating 100 neurons for 5 time-steps. On each time-step of stimulation input, random noise was applied so that 30% of the input was randomly deactivated. Prior to each trial, baseline activity was simulated by providing the network with a new random input initialization vector such that a random 7.5% of the neurons were active. The synaptic weights were all initialized to 0.4 prior to learning, and the synaptic weight modification rate was 0.01 per time-step. The parameter α was set to 0.8. Each simulation took approximately 28 minutes to train on a 2.66 GHz quad-core Intel Xeon as implemented in single-threaded Java code (source code is available from http://pakl.net/code/hc/).

Assessing predictive similarity. To quantify how much discounting will occur for a delayed reward in this model, it is necessary to measure the similarity between the pattern in the CA3 network at the time of the decision to the pattern of hippocampal activity which represents the delayed rewarded stimulus. According to our hypothesis, the similarity will decrease with in-

creased delay between the decision time and the time at which the rewarded stimulus is predicted to occur. This comparison is quantified by treating network activity as a vector and using cosine similarity (i.e., the normalized dot product) between pairs of network state (i.e., neuronal firing) vectors,

$$sim(z_1, z_2) = \frac{\overrightarrow{z_1} \cdot \overrightarrow{z_2}}{|\overrightarrow{z_1}||\overrightarrow{z_2}|} \tag{5}$$

which returns a similarity value ranging from zero (i.e., orthogonal vectors) to 1 (i.e., identical vectors).

The similarity of the current pattern to the discounted future pattern will, through connections to the ventral striatum via CA1 (Kelley & Domesick, 1982), partially re-activate ensembles and reward predictions for the future stimuli (Lansink et al., 2009; Pennartz et al., 2004).

Results

Our hypothesis is that hyperbolic reward discounting occurs because at the time of decision-making, the organism makes a comparison between (1) reward predictions associated with the immediate stimulus and (2) reward predictions associated with a *discounted* representation of the future stimulus. Thus, the amount of discounting is hypothesized to depend directly on the extent to which the hippocampal representation of the present is similar to the hippocampal representation of the future, anticipated stimulus when

it is activated (and when reward was delivered). To test this hypothesis, we quantified the degree to which the representation for the present code resembled the representation when the future stimulus was present – i.e., its predictive similarity.

To provide a visual sense of predictive similarity in the hippocampal component of the model, Figure 1 shows the firing over time of the simulated neurons that were used as input neurons (i.e., they were clamped as part of the stimulation). By examining this figure, one obtains an explicit sense of predictive similarity: the activity within the dotted box shows that neurons that were normally clamped for the third pattern in the sequence have started firing during the second pattern. Although the figure only shows the clamped neurons, it is important to note that these neurons are effectively restricted in terms of how far into the past they can shift because of inhibition and competition in the network. In contrast, non-clamped neurons in the rest of the network are "free" to shift earlier in time and lead to predictive similarity that extends over multiple stimuli, not just temporally adjacent stimuli.

The cosine similar between the firing pattern one quarter of the way into the sequence and each subsequent pattern in the sequence was computed. (The initial and final quarters of the sequences were excluded from the analysis to ensure that measures reflected on-line function, and to avoid biases that might occur due to end effects.) The result for 50 simulated networks is shown graphically in Figure 2A. As can be seen in that figure, the similarity of

the code for the present possesses decreasing similarity to codes for patterns farther into the future, in a manner that approximates hyperbolic discounting. Indeed, a hyperbolic function is a good fit to this measure of predictive similarity ($y = 1/(1+0.29t)$, SSE = 0.0084). The fitted hyperbolic discount curve is plotted in Figure 2B. These results suggest that hippocampal predictive similarity may be a suitable mechanism for understanding reward delay discounting.

As mentioned in the introduction, the predominant account of hippocampal contributions to decision-making is the "mental simulation" account according to which future reward payoffs are evaluated by imagining future events. For example, studies show that when rats deliberate about whether to go left or right in a T-maze, neurons in their hippocampus fire as if they were exploring the two options in the maze (Johnson et al., 2007; Johnson & Redish, 2007). Although it is clear that this activation could be used to select actions based on delayed outcomes, it is not immediately obvious whether or how this mechanism might lead to reward delay discounting. One hypothesis is that the mental simulation account is compatible with present model. Indeed, we can test this hypothesis: the present model can be run in sort of "mental simulation" mode by providing it with an input pattern at the beginning of a sequence, and then allowing it to activate subsequent representations without any further input (August & Levy, 1999). A test of the mental simulation account within the context of this model then involves quantifying the extent to which the recalled representations resemble

the original input activations and thereby activate the associated reward predictions. That is, we can operationalize the mental simulation account of reward discounting is as follows: "Reward discounting occurs because of decreasing robustness (i.e., decreasing similarity) of recalled activity to the original activity generated by the stimuli as the simulation proceeds farther into the future[1].

We measured how similar the recalled patterns were to the patterns during training during time-steps 5 through 50 (see Figure 3). Counter to the predictions of the mental simulation account, the maximal similarity between recall and training patterns did not decrease but rather increased slightly over the course of recall in the 50 simulated networks (mean slope of 0.0005 was greater than 0, $t(49) > 2.78$, $p < 0.0075$). Thus, the robustness of the recalled patterns does not decrease across the sequence, indicating that the present model makes predictions that are not compatible with the mental simulation account.

Discussion

We took an existing computational model of hippocampal function and used it develop a novel, mechanistic explanation for reward discounting. This ex-

[1]Note that this is a different measure from a decrease in similarity to the current activity pattern, which is what we have quantified above as "predictive similarity". To be precise, indicate a stimulated pattern by a capital letter but a recalled pattern by a capital letter prime. Then predictive similarity is about the similarity between A, and B, C, D, etc., whereas mental simulation is about the similarity between A and A', B and B', C and C', etc.)

planation suggests how projections from hippocampus to the striatum might be part of a network of brain systems that explain observed patterns in decision-making behavior about rewards (see Figure 4). An analysis of simulation results showed that, in the model, reward delay discounting arises when a recurrent network re-encodes input stimuli so that their similarity to future stimuli is increased. At decision time, the magnitude of this "predictive similarity" approximates a hyperbolic function. That is, the similarity of hippocampal output to future patterns decays with increasing temporal delay to the rewarded stimulus. Because the hippocampus has a strong functional projection to the ventral striatum, it follows that activity generated in the reward system for anticipated stimuli might also reflect this decay. This suggests that predictive similarity recoding could be an explanation of reward discounting and its particular functional form.

Although we cannot make any strong claims about whether the mental simulation account plays a role in reward delay discounting, the present model seems to be incompatible with that account. Mental simulation proposes that future rewards influence decisions because at the time of a decision, the hippocampus simulates paths to rewarded stimuli. When we operationalized this in our model by allowing the CA3 network to perform sequence recall, we found that patterns were recalled with equal robustness throughout the sequence. Thus, although it is clear that mental simulation occurs and may be useful for cognition, the present simulation results argue against a role for mental simulation in reward discounting. Future experimental and

modeling work would be helpful to provide insight into understanding contributions of the mental simulation and predictive similarity mechanisms.

Interestingly, the mental simulation account also appears to be at odds with observed behavioral data: Because mental simulation should take time proportional to the delay until the (mentally simulated) reward, the mental simulation account should predict that decisions involving delayed stimuli should take longer to make, as longer sequences are simulated. However, this is not the case according to the data in Peters and Büchel. They examined reaction time by condition and by (undiscounted) value, and found that reaction time was independent of condition (see their Figure 2C).

The equally-robust recall throughout the sequence likely occurred because this network, as do others with time-spanning associative learning rules, forms transient attractors during learning (Liljenström & Wu, 1995; Sompolinsky & Kanter, 1986; Wu & Liljenstrom, 1994). Because the input patterns were orthogonal and all received the same number of training trials, the transient attractors during the sequence all had a similar shape. This brings up an interesting possibility for future work with this model which may provide explanations for the results of Peters and Büchel (2010). As mentioned above, they found that participants attending future events exhibited a decrease in reward discounting, and that this appeared to be mediated by increased functional connectivity between the prefrontal cortex and hippocampus. This increase in functional connectivity may influence the effective salience of an intervening stimulus (e.g, which could be simulated

by repeated exposures to that stimulus during learning or increased number of activated inputs for that stimulus). Each of these manipulations will have an effect on the form of the transient attractors, and this might lead to an increase in similarity for nearby patterns and a reduction of reward discounting – capturing the result found by Peters and Büchel (2010).

Another possible mechanism that may also explain their result is that attentional control regions might increase the gain of the hippocampal influence on the striatal reward-predicting network. This input might regulate the degree to which the hippocampal output contributes to striatal activity (e.g., Figure 4C). Such a projection could allow attention to enhance the activity and retrieval of both spatial and non-spatial representations in the hippocampus that are associated with reward (Muzzio et al., 2009). Thus there are at least two mechanisms that should be explored to explain the modulatory effects of attention on reward delay discounting.

This article focused on hippocampal-striatal interactions because of the well-established recurrent architecture of the hippocampus and the well-known role of the striatum in reward processing. However it is possible that recurrent networks in other brain regions (e.g., perhaps in the prefrontal cortex) could influence the reward predictions and thus implement reward discounting. Because reward delay discounting appears to operate on many different timescales ranging down to the level of milliseconds for saccadic eye movements (Shadmehr et al., 2010), it seems likely that diverse brain mechanisms underly reward delay discounting.

Acknowledgements

I thank William B Levy, Joe Monaco and Sean Polyn for comments on an earlier version of this manuscript. This work was funded by NIH grant R01-DA013165 to Steve Yantis.

References

Addis, D. R., Cheng, T., P Roberts, R., & Schacter, D. L. (2010). Hippocampal contributions to the episodic simulation of specific and general future events. *Hippocampus*, .

August, D. A., & Levy, W. B. (1999). Temporal sequence compression by an integrate-and-fire model of hippocampal area ca3. *J Comput Neurosci*, *6*, 71–90.

Bamford, N. S., Zhang, H., Schmitz, Y., Wu, N.-P., Cepeda, C., Levine, M. S., Schmauss, C., Zakharenko, S. S., Zablow, L., & Sulzer, D. (2004). Heterosynaptic dopamine neurotransmission selects sets of corticostriatal terminals. *Neuron*, *42*, 653–63.

Bi, G. Q., & Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J Neurosci*, *18*, 10464–72.

Cheung, T. H. C., & Cardinal, R. N. (2005). Hippocampal lesions facili-

tate instrumental learning with delayed reinforcement but induce impulsive choice in rats. *BMC Neurosci*, *6*, 36.

Figner, B., Knoch, D., Johnson, E. J., Krosch, A. R., Lisanby, S. H., Fehr, E., & Weber, E. U. (2010). Lateral prefrontal cortex and self-control in intertemporal choice. *Nat Neurosci*, *13*, 538–9.

Gamboz, N., Brandimonte, M. A., & De Vito, S. (2010). The role of past in the simulation of autobiographical future episodes. *Exp Psychol*, *57*, 419–28.

Gupta, R., Duff, M., Denburg, N., Cohen, N., Bechara, A., & Tranel, D. (2009). Declarative memory is critical for sustained advantageous complex decision-making. *Neuropsychologia*, *47*, 1686–1693.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmpfc valuation system. *Science*, *324*, 646–8.

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proc Natl Acad Sci U S A*, *104*, 1726–31.

Hollerman, J. R., Tremblay, L., & Schultz, W. (1998). Influence of reward expectation on behavior-related neuronal activity in primate striatum. *J Neurophysiol*, *80*, 947–63.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw*, *15*, 535–47.

Johnson, A., van der Meer, M. A. A., & Redish, A. D. (2007). Integrating hippocampus and striatum in decision-making. *Curr Opin Neurobiol*, *17*, 692–7.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *J Neurosci*, *27*, 12176–89.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nat Neurosci*, *10*, 1625–33.

Kable, J. W., & Glimcher, P. W. (2010). An "as soon as possible" effect in human intertemporal decision making: behavioral evidence and neural mechanisms. *J Neurophysiol*, *103*, 2513–31.

Kacelnik, A. (1997). Normative and descriptive models of decision making: time discounting and risk sensitivity. *Ciba Found Symp*, *208*, 51–67; discussion 67–70.

Kelley, A., & Domesick, V. (1982). The distribution of the projection from the hippocampal formation to the nucleus accumbens in the rat: An anterograde and retrograde-horseradish peroxidase study. *Neuroscience*, *7*, 2321–2335.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., & Pennartz, C. M. A. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol*, *7*, e1000173.

Levy, W. (1989). A computational approach to hippocampal function. *Computational models of learning in simple neural systems*, *23*, 243–305.

Levy, W. B. (1996). A sequence predicting ca3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, *6*, 579–90.

Levy, W. B., Sanyal, A., Rodriguez, P., Sullivan, D. W., & Wu, X. B. (2005). The formation of neural codes in the hippocampus: trace conditioning as a prototypical paradigm for studying the random recoding hypothesis. *Biol Cybern*, *92*, 409–26.

Levy, W. B., & Steward, O. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, *8*, 791–7.

Liljenström, H., & Wu, X. B. (1995). Noise-enhanced performance in a cortical associative memory model. *Int J Neural Syst*, *6*, 19–29.

Mariano, T. Y., Bannerman, D. M., McHugh, S. B., Preston, T. J., Rudebeck, P. H., Rudebeck, S. R., Rawlins, J. N. P., Walton, M. E., Rushworth, M. F. S., Baxter, M. G., & Campbell, T. G. (2009). Impulsive choice in

hippocampal but not orbitofrontal cortex-lesioned rats on a nonspatial decision-making maze task. *Eur J Neurosci*, *30*, 472–84.

Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, *275*, 213–5.

Mazur, J. (1987). An adjusting procedure for studying delayed reinforcement. *Quantitative Analyses of Behavior: The Effects of Delay and of Intervening Events on Reinforcement Value*, *5*, 55.

Mazur, J. E., & Biondi, D. R. (2009). Delay-amount tradeoffs in choices by pigeons and rats: hyperbolic versus exponential discounting. *J Exp Anal Behav*, *91*, 197–211.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *J Neurosci*, *27*, 5796–804.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*, 503–7.

Mehta, M., Barnes, C., & McNaughton, B. (1997). Experience-dependent, asymmetric expansion of hippocampal place fields. *Proceedings of the National Academy of Sciences of the United States of America*, *94*, 8918.

Mehta, M., Quirk, M., & Wilson, M. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, *25*, 707–715.

Mitman, K., Laurent, P., & Levy, W. (2003). Defining time in a minimal hippocampal ca3 model by matching time-span of associative synaptic modification and input pattern duration. In *Neural Networks, 2003. Proceedings of the International Joint Conference on* (pp. 1631–1636). IEEE volume 3.

Muzzio, I. A., Levita, L., Kulkarni, J., Monaco, J., Kentros, C., Stead, M., Abbott, L. F., & Kandel, E. R. (2009). Attention enhances the retrieval and stability of visuospatial and olfactory representations in the dorsal hippocampus. *PLoS Biol*, *7*, e1000140.

Pennartz, C. M. A., Lee, E., Verheul, J., Lipa, P., Barnes, C. A., & McNaughton, B. L. (2004). The ventral striatum in off-line processing: ensemble reactivation during sleep and modulation by hippocampal ripples. *J Neurosci*, *24*, 6446–56.

Peters, J., & Büchel, C. (2010). Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron*, *66*, 138–48.

Rawlins, J. N., Feldon, J., & Butt, S. (1985). The effects of delaying reward on choice preference in rats with hippocampal or selective septal lesions. *Behav Brain Res*, *15*, 191–203.

Reynolds, J. N. J., & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw*, *15*, 507–21.

Schacter, D. L., & Addis, D. R. (2009). On the nature of medial temporal lobe contributions to the constructive simulation of future events. *Philos Trans R Soc Lond B Biol Sci*, *364*, 1245–53.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–9.

Shadmehr, R., Orban de Xivry, J. J., Xu-Wilson, M., & Shih, T.-Y. (2010). Temporal discounting of reward and the cost of time in motor control. *J Neurosci*, *30*, 10507–16.

Sompolinsky, & Kanter, I. (1986). Temporal association in asymmetric neural networks. *Phys Rev Lett*, *57*, 2861–2864.

Suri, R. E., & Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Exp Brain Res*, *121*, 350–4.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: an introduction. *IEEE Trans Neural Netw*, *9*, 1054.

Wu, X., & Liljenstrom, H. (1994). Regulating the nonlinear dynamics of olfactory cortex. *Network: Computation in Neural Systems*, *5*, 47–60.

Figures

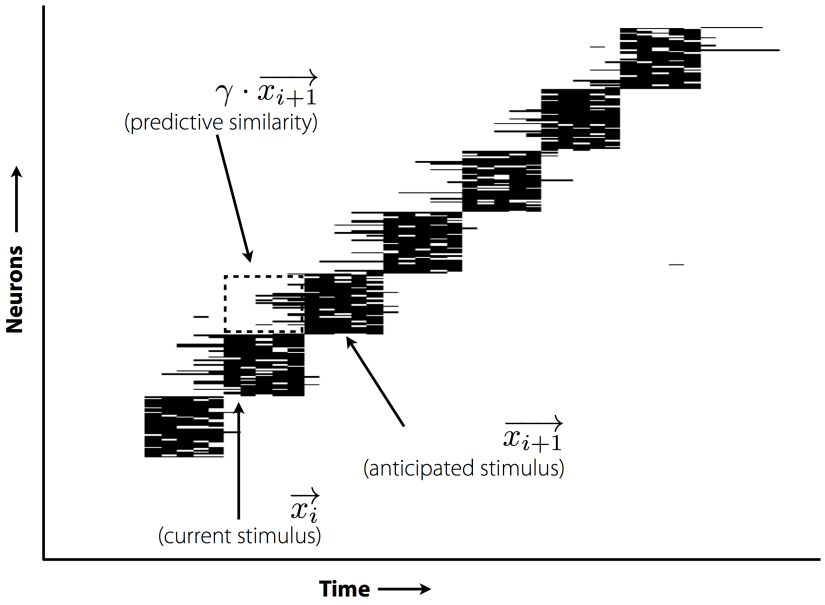

Figure 1: **Ealier-shifting of firing times giving rise to 'predictive similarity'.** Reward discounting is hypothesized to emerge when neurons in region CA3 in the hippocampus shift in time to become associated with predictive stimuli. Earlier-shifting is illustrated most readily in the field of neurons that were activated by the external stimuli. The dotted line shows firing from the future pattern that has come to be activated as part of the firing associated with an earlier stimulus. Note that other neurons in the network may shift much earlier because they are not effectively "anchored" in time by the external clamping stimulation. Similarity arising from this recording can be quantified using a similarity measure, as discussed in the text.

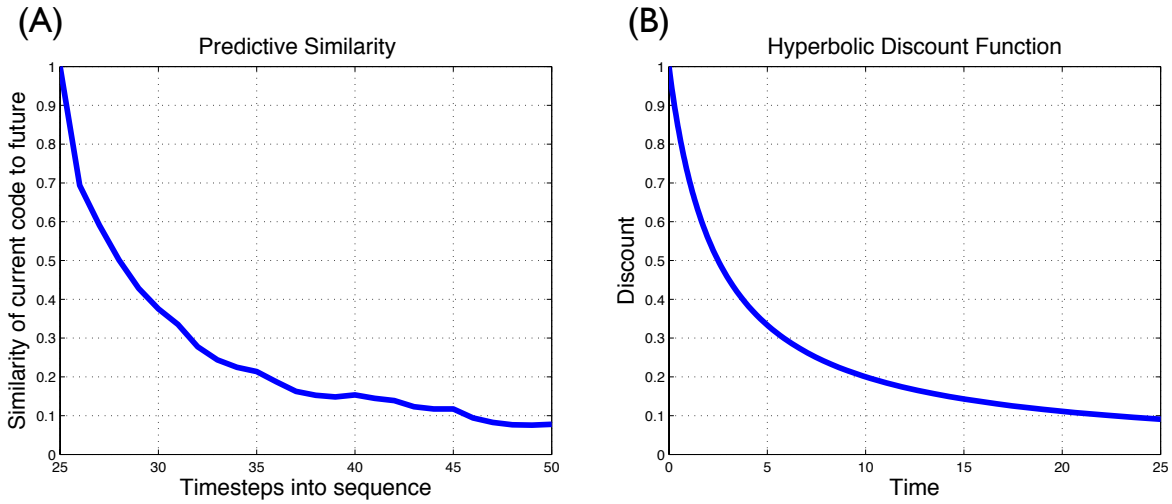

Figure 2: **Similarity between simulated hippocampal activity pattern and subsequent patterns proposed as a neural basis for reward discounting.** The plot on the left shows how the similarity between the activity pattern at time-step 25 and subsequent patterns decrease as they are separated in time (average of N=50 simulations). The plot on the right shows the fitted hyperbolic function for comparison.

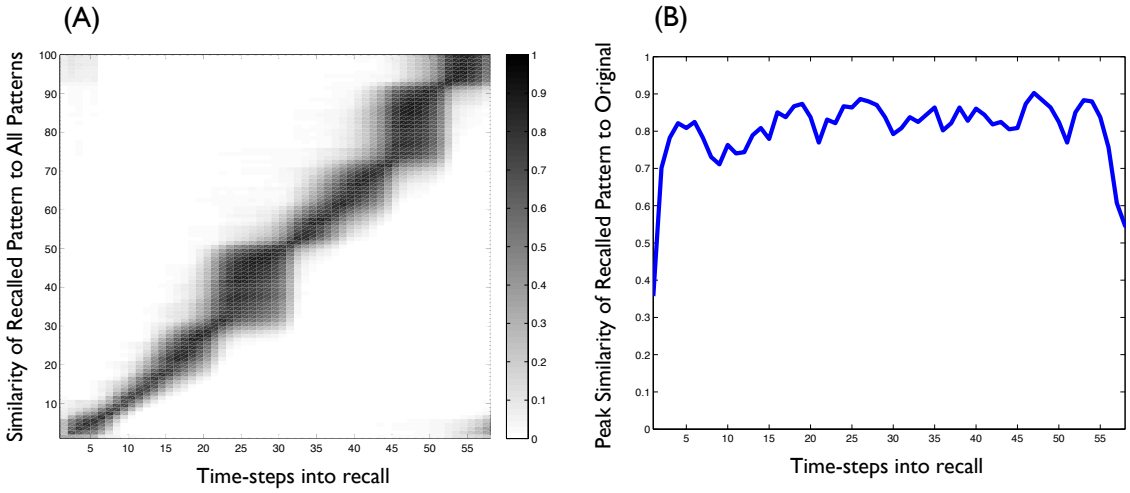

Figure 3: **Operationalizing Mental Simulation.** The present framework can be tested for compatibility with the mental simulation account by examining the activity patterns recalled in sequence. According to the mental simulation account of reward delay discounting, patterns decrease in similarity to their original patterns. However, that hypothesis is not supported by the present model: **(A)** On the left is plotted the similarity between each recalled pattern (x-axis) and its original pattern (y-axis). **(B)** The right panel shows that the maximum amount of similarity on each time-step is approximately constant throughout the entire free-recall simulation sequence, with a mean value of 0.81. That is, each mentally-simulated stimulus is activated equally robustly as simulation proceeds until the end of the sequence. Thus, because the fidelity of the simulated experience did not decrease as the sequence proceeded, these results suggest that the present account is not compatible with the mental simulation account.

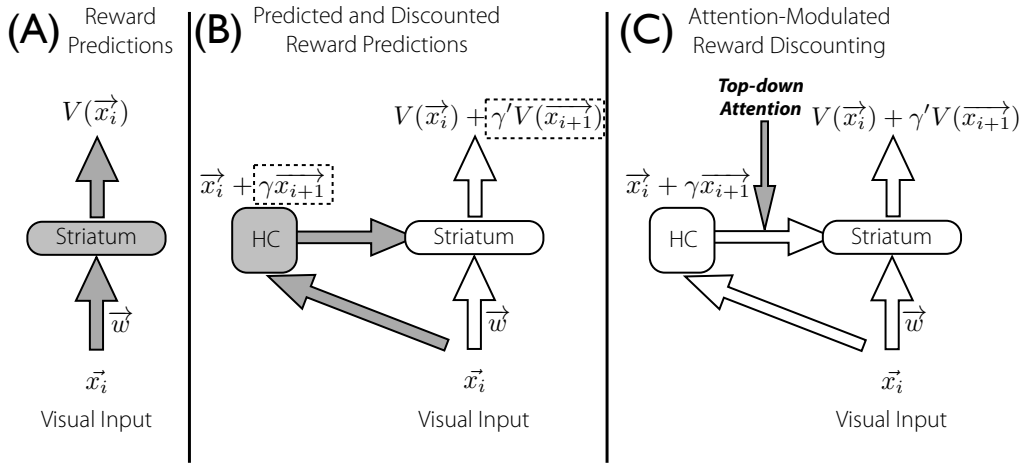

Figure 4: **Proposed system interactions to explain reward discounting.** From left to right, the panels show the construction of the complete model described in the text. **(A)** In typical reward-prediction network implementations using function approximators, a state representation x (e.g., visual stimulus) is presented to a neural network, is processed through a set of putatively cortico-striatal weights, and generates a reward prediction V(x). In Reinforcement Learning, this value can be used for optimal decision making or action selection. **(B)** The hippocampus (HC) forms a neural representation through learning that has partial predictive similarity to future states. Thus, hippocampal input to the striatum partially activates reward predictions for future stimuli, augmenting the reward prediction to include a discounted portion of the future reward. **(C)** The hippocampus' contribution can be modulated by top-down attention, which might increase the gain of the hippocampo-striatal projection. This would increase the contribution of predictive similarity to the final reward prediction output.