Evidence for two distinct mechanisms directing gaze in natural scenes

Michael Mackay
Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA, &
Now at the School of Clinical Medicine,
University of Cambridge, England

Moran Cerf
Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA, &
Now at New York University, New York, NY, USA

Christof Koch
Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA, &
Now at the Allen Institute for Brain Science,
Seattle, WA, USA

Various models have been proposed to explain the interplay between bottom-up and top-down mechanisms in driving saccades rapidly to one or a few isolated targets. We investigate this relationship using eye-tracking data from subjects viewing natural scenes to test attentional allocation to high-level objects within a mathematical decision-making framework. We show the existence of two distinct types of bottom-up saliency to objects within a visual scene, which disappear within a few fixations, and modification of this saliency by top-down influences. Our analysis reveals a subpopulation of early saccades, which are capable of accurately fixating salient targets after prior fixation within the same image. These data can be described quantitatively in terms of bottom-up saliency, including an explicit face channel, weighted by top-down influences, determining the mean rate of rise of a decision-making model to a threshold that triggers a saccade. These results are compatible with a rapid subcortical pathway generating accurate saccades to salient targets after analysis by cortical mechanisms.

Keywords: eye-tracking, natural scenes, saliency, saccadic latency, drift diffusion model

Citation: Mackay, M., Cerf, M., & Koch, C. (2012). Evidence for two distinct mechanisms directing gaze in natural scenes. *Journal of Vision, 12*(4):9, 1–12, http://www.journalofvision.org/content/12/4/9, doi:10.1167/12.4.9.

Introduction

In order to decide where to look in a scene, observers need information on what is present in their visual field. It is generally believed that gaze is controlled by bottom-up, task-independent strategies in combination with top-down, goal-driven information (Itti, Koch et al., 1998; Oliva, Torralba et al., 2003; Reddi & Carpenter, 2000; Yarbus, 1967). Scene-specific information, such as gist (Biederman, 1987; Fei-Fei, Iyer, Koch & Perona, 2007; Torralba, Oliva et al., 2006), takes some time to reach higher regions (Bar, Kassam et al., 2006), partially explaining the long latency of many saccades (Leach & Carpenter, 2001).

However, saccadic latency is still longer and more variable than would be expected based on the latencies of visual processing. This is thought to be as a result of the process of saccadic decision.

The LATER model (Carpenter & Williams, 1995; Reddi & Carpenter, 2000) is a commonly used race-to-threshold model of saccadic decision (see also Smith & Ratcliff, 2004 and Ratcliff & McKoon, 2008). It proposes that a detection signal S rises linearly from a starting value, S_0, at a rate r until some threshold level, S_T, is reached, at which point in time, T, a saccade is triggered. S_0 represents any initial bias with $S_0 = 0$ implying none. If r varies randomly from saccade to saccade with a Gaussian distribution, then the result will be a latency histogram with a tail skewed to longer latencies as is commonly observed. More specifically, the saccadic latency distribution can be reflected as a straight line when plotted cumulatively on a *probit* ordinate and *reciprocal* abscissa, a *reciprobit* plot (Figures 1a and 1b).

With large data sets, however, we see more early responses than a Gaussian distribution for the rate of rise, r, would predict. These early responses can be fitted by a separate trend line that intersects the T = infinity axis at 50%. This trend line is of shallower slope than the main distribution intersecting it, and is more pronounced when the target is expected or there is a high degree of urgency in the task (Reddi & Carpenter, 2000). These early responses may include express saccades. However,

doi: 10.1167/12.4.9 Received March 9, 2011; published April 17, 2012 ISSN 1534-7362 © ARVO

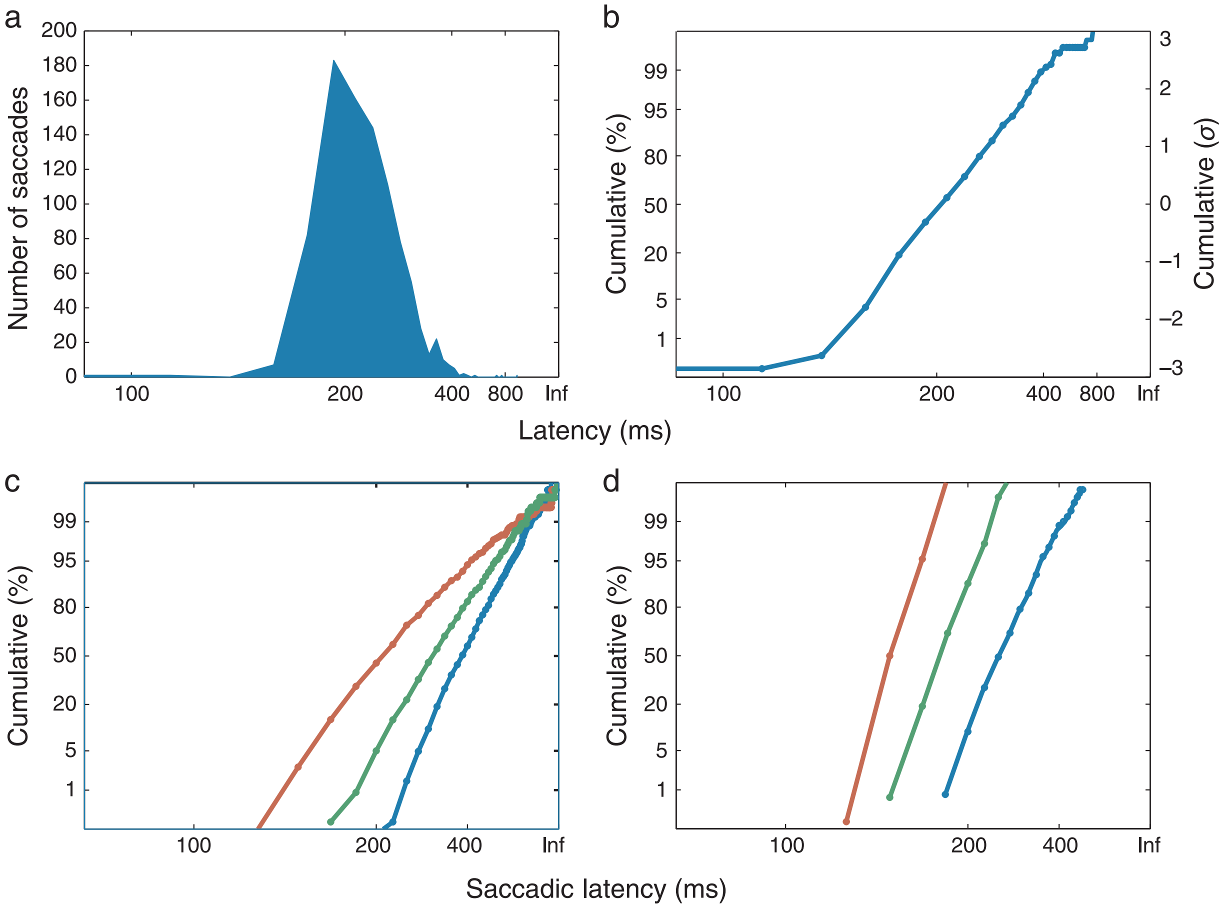

Figure 1. Transformations of the LATER model. (a, b) Data are taken from first fixations from a single subject. (a) When a latency histogram is plotted on an inverse, 1/*T,* or reciprocal abscissa, the distribution becomes more symmetrical. (b) The histogram data plotted on a probit ordinate and reciprocal abscissa produce a straight line, a reciprobit plot. The ordinate is in effect a linear axis where the unit is standard deviations taken from the mean at 0. (c, d) Data generated by simulation of the LATER model. (c) Mean and variance remain constant; S_T is set to 100. Increasing the prior probability, S_0, of fixating on some object from unbiased at 0 (blue line), to 25 (green) and 50 (red), causes clockwise swivel about the *y* intercept, reducing the gradient and shifting the saccadic distribution to earlier latencies. (d) Prior and threshold probability and variance are held constant, while the mean rate of rise (μ) is increased from 5 (blue) to 7.5 (green) and 10 (red). This leads again to faster saccadic times without shallower curves due to parallel leftward shift of the curves with increasing μ.

express saccades form a distinctly bimodal distribution and these early responses are only apparent on a reciprobit plot, unlike express saccades (Carpenter, 2001; Fischer & Ramsperger, 1986). Early saccades are thought to represent relatively automatic responses, mediated by a subcortical structure such as the superior colliculus (Carpenter, 1994; Schiller, Sandell et al., 1987).

Functionally, a decision making model, (e.g. the LATER model) represents an ideal Bayesian decision-making process (Reddi, Asrress, & Carpenter, 2003). The decision signal S represents the log likelihood of a hypothesis ("there's something to look at here") being correct at any given time. The initial value S_0 represents the logarithm of prior probability. S rises linearly at a rate r, with incoming confirmatory sensory information until it reaches a threshold S_T. This threshold reflects the probability that justifies the initiation of a saccade. Thus, the reciprocal of latency $1/T$ can be described by the following equation:

$$\frac{1}{T} = \frac{r}{S_T - S_0} \qquad (1)$$

As r is a Gaussian random variable with mean μ and standard deviation σ, the distribution of $1/T$ will also vary in a Gaussian manner with a mean of $\mu/(S_T - S_0)$ and variance of $\sigma^2/(S_T - S_0)^2$. Any proportional change in both S_T and S_0 will have no effect on the distribution. Thus, they can be combined as ΔS. Equally, the distribution would be unchanged by a proportional change in μ, σ, and ΔS. As such, the system has two degrees of freedom; determination of any two of these three parameters defines the system completely.

We can describe the transformations of the reciprobit plot we expect to see on manipulation of these parameters.

As this is a Gaussian distribution, the intercept of the curve with the 50% axis defines both the median and mean and is given by $1/T = \mu/\Delta S$. The gradient is given by $\Delta S/\sigma$ and the y intercept (the intercept with T = infinity) is defined by μ/σ. This intercept corresponds to the probability that no saccade will be generated in finite time (i.e., $r \leq 0$).

If we change ΔS, we change the gradient and the median but not the y intercept, thus causing a swivel of the curve around the intercept with the T = *infinity* axis. Increases in prior probability, S_0 (Carpenter & Williams, 1995), and decreases in threshold probability, S_T (Reddi & Carpenter, 2000), reduce ΔS and thus reduce the gradient causing a clockwise swivel around the intercept. Decreases in S_0 and increases in S_T have the opposite effect (Figure 1c). If we alter the *mean rate of rise*, μ, we move both the y intercept and the intercept with the horizontal 50% axis (the median), but the gradient remains unchanged. Thus, we shift the curve in parallel (Reddi et al., 2003). An increase in μ corresponds to a leftward shift in the curve, a decrease to a rightward shift (Figure 1d).

We can consider a decision system to be constructed of many parallel "units," each functioning independently according to the properties of a LATER model. These units each represent their own "hypothesis" or visual target to saccade to and form a race-to-threshold model of decision making. Such competitive racing can be seen in a precedence task, whereby a target appears in the subject's field of view, followed after a small delay, δ, by a second, competing target (Leach & Carpenter, 2001). Where δ is large, the first target is usually fixated first. As δ becomes smaller, the first target is fixated proportionally less, and at $\delta = 0$ ms, the fixations are split 50/50 to each target. As well as demonstrating competitive racing, this paradigm also demonstrates independent randomness between units. If the randomness were correlated, the target that appears first would always win.

One corollary of the LATER model is that this independent, random variation in rate of rise effectively leads to a randomization of choice (Carpenter, 1999). This may not seem ideal in a decision mechanism, but in many situations a winning strategy is to make your behavior as unpredictable as possible; a predictable behavior is one that can be countered (von Neumann & Morgenstern, 1944). However, that is not to say that an organism functions in a truly random manner; a decision mechanism should ideally take into account the expected utility of any decision it makes (Good, 1952).

Additionally, when multiple saccades are made to a stationary scene, later saccades can take advantage of visual information acquired during earlier fixations (Kotowicz, Rutishauser, & Koch, 2010). As this requires the storage of information across saccades, this is likely to involve additional neural mechanisms (Khayat, Spekreijse et al., 2004).

Analysis of fixations in natural scenes provides a means to isolate these features of decision models by using the properties of the LATER model, combined with novel fixation-by-fixation analysis of saccadic latencies and inter-saccadic intervals. We hypothesized that there would be increases in prior probability, as seen by swivel in the reciprobit plots, representing the utilization of visual information gathered earlier during scene viewing. This swivel could then be augmented by changes in the mean rate of rise of LATER units as a means of biasing decision.

We provide evidence that supports the existing interpretation of the LATER model. We extend this model by proposing that incoming sensory information is weighted by bottom-up and top-down mechanisms to drive changes in the mean rate of rise of individual LATER units. As such, we still have a random decision mechanism, but the mean rate of rise of a unit effectively represents the bias of that unit to selection, which we propose correlates to a measure of expected utility of that decision. We investigate the nature of early saccades in natural scenes and show that these saccades do represent automatic responses, as seen in evoked saccadic tasks, but can later take advantage of prior visual information to target semantically meaningful objects and faces. We propose that this reflects a mechanism to rapidly direct gaze to salient objects during scene scanning without needing to wait for costly cortical analysis of each new retinal image.

Methods

Task

Nineteen subjects viewed 680 photos (1024 × 768 pixels) depicting natural scenes for 2 s each under two distinct instructions using the methods described in Cerf, Frady, and Koch (2009). These images were viewed under "free-viewing" conditions, whereby subjects had to answer the question "how interesting was the previous image?" using a scale of 1–9 (9 being the most interesting). Subjects were not instructed to look at anything in particular; their only task was to rate the entire image. Subjects do this task in a manner that is consistent and stable across months and across subjects (Cerf, Cleary et al., 2007). A further 200 images were viewed under "search" conditions, whereby observers first viewed a target (a face or an object) for 600 ms and then had to inspect an image that either did or did not contain that target for 2 s. Subsequently, they had to judge whether or not the target appeared in the previous image. Half of the images contained the target, while half of the targets were faces.

Visual stimuli

Subjects saw a total of 880 images in 5 experiments. In 3 experiments totaling 600 images, 406 face-containing images were analyzed, of which 151 images were unique;

200 of these images were viewed in a fourth experiment under “search” conditions. A fifth experiment contained 80 images with text (e.g., shop or street signs, movie marquee) taken from the Internet. The face images were photographed in indoor and outdoor environments. The images included people of various skin colors, ages, and postural positions. A few images had face-like objects (e.g., smiley T-shirt, animal faces, masks, faces carved in stone). Some of the images contained objects such as a colorful Rubik’s cube, a toy fire truck, plastic banana, and other visually salient objects. These objects competed with faces and text elements for gaze. The average sizes of the faces/text/objects were 4.0% ± 1.8% of the images by area. The area of the faces/text/objects was calculated from a hand-drawn region covering the entire region of interest. The images and subjects’ scan paths are taken from www.fifadb.com (Cerf et al., 2009). Image order within each experiment was randomized throughout the experiment.

Data collection

The data were acquired at 1000 Hz using an infrared Eyelink 1000 eye-tracking device (SR Research, Osgoode, Canada) and a chin rest. The 1000-Hz samples acquired by the eye tracker allow for real-time calculation of velocity, based on the x and y positions at any given millisecond. These are the absolute velocities measured as the Euclidean sum of x and y components. The EyeLink 1000 parser computes velocity by use of a 9-sample moving filter. For each data sample, the parser computes instantaneous velocity and acceleration and compares these to the velocity and acceleration thresholds. If either is above threshold, a saccade signal is generated. The parser checks that the saccade signal is on or off for a critical time before deciding that a saccade has begun or ended (Cerf, Harel et al., 2008). Following a calibration process, subjects initiated the experiment. Prior to each stimulus presentation, the subjects were instructed to look at a black fixation cross at the center of the screen. If the calculated gaze position was not at the center of the screen, the calibration process was repeated to ensure that position was consistent throughout the experiment. Images were presented on a CRT2 screen (120 Hz), using Matlab’s Psychophysics and the Eyelink toolbox extension. Stimulus luminance was linear in pixel values. The distance between the screen and the subject was 80 cm, giving a total visual angle for each image of 28° × 21°. Subjects used a chin rest to stabilize their head. Eye movement data were acquired from the right eye alone. All subjects had normal or corrected-to-normal eyesight. All subjects were naive to the purpose of the experiment. The experiment was undertaken with the understanding and written consent of each subject. All experimental procedures were approved by Caltech’s Institutional Review Board.

Fixations were determined by the built-in software of the eye-tracking system. The “initial fixation” is always to the fixation-centered cross and is not counted as part of the ordered sequence of fixations. On image presentation, the first saccade made ends with the first fixation.

To compute chance level of performance of fixations, we calculated the fraction of all subjects’ fixations from all other images that fall into the ROI of each particular image. This takes into account the varying size and locations of the ROI in all images (as these factors both influence how likely a certain region is to be fixated on by chance) and the spatial bias of photographer and observer.

Results

In order to test how subjects respond to natural scenes, we analyzed the latencies of saccades that land on various targets during free-viewing and search tasks containing faces, objects, and text elements. Saccadic latencies were separated out on the basis of fixation number to an image. For first fixations, saccadic latency was calculated as the time from image onset to initiation of the first saccade. For the following fixations, saccadic latency was calculated as the time between the end of the previous saccade and the initiation of the current saccade.

In line with previous results, we found a latency histogram with a tail skewed to longer latencies for saccades evoked by image onset (Figure 2a). When plotted on a reciprobit, it forms a characteristic straight line (Figure 2c). All following saccades are then “spontaneous” saccades and breaking these saccades down on an individual fixation basis produces typically skewed latency histograms (Figure 2b). When these spontaneous saccades are plotted on a reciprobit, a curved line emerges, with a large and evident population of early saccades, in line with previous studies (Figure 2d; Roos, Calandrini et al., 2008).

We found that the vast number of fixations fall on one of two distinct manifolds in reciprobit plots, an early distribution and a main distribution (Figures 2c and 2d). The cutoff between the two occurs at 122.1 ± 24.7 ms (mean ± std) for the first fixation and at 108.9 ± 10.2 ms for subsequent fixations (averaged over all subjects and images). The fraction of early saccades for the first saccades was 3.0 ± 2.4% and 4.2 ± 1.7% for all subsequent saccades.

Proportion of early saccades landing on faces and text

The clear distinction between the main distribution of saccades (MS) and the early distribution of saccades (ES) makes the two populations amenable to investigation. While it was shown before that first saccades normally

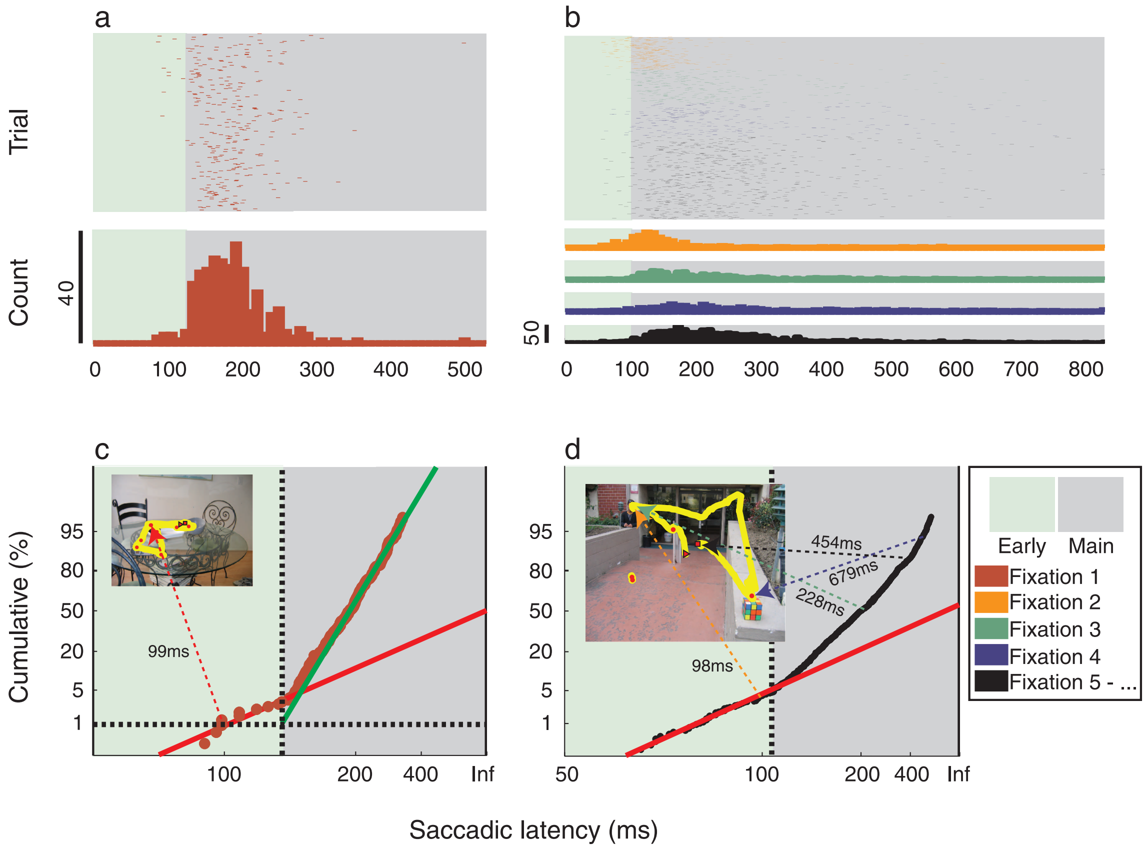

Figure 2. Examples of saccadic latency fits of subjects. (a) Saccadic latencies of the first fixations of one arbitrarily chosen subject viewing 344 images. The latencies are color coded by a shaded rectangle, reflecting the breakdown between the categories of "early" and "main" saccades. (b) Second and subsequent fixations. The distributions of 2nd fixation saccades show a decrease in latencies compared to the first saccade (in (a)). Subsequent fixation distributions then show a progressive increase in latency compared to the 2nd. (c) First fixations from (a) and the corresponding fit to a Gaussian on a cumulative plot with reciprocal abscissa. Two distributions are seen—early ("maverick"; 3.5% of all fixations) and main distribution saccades—separated at 130 ms and fit by red and green lines, respectively. The early saccades were fitted such that the *y*-axis intercepts infinity at 50%. An example of the first fixation latency and location for the subject and one image are shown, with the saccade latency being 99 ms. (d) Second and subsequent fixations from (b). The cutoff from early to main is at 105 ms. Second and higher fixations show an increased proportion of early saccades that make up 4.9% for 2nd and higher saccades. An example image viewed by this subject and his corresponding 2nd to 5th saccadic latencies and locations are shown.

fixate faces (Cerf et al., 2009), we hypothesized that this requires information not available to the neural networks driving early saccades, and therefore, early saccades would show a lower proportion of fixation on faces. We normalized the latencies, such that a latency of 0 ms represents the cutoff between the early and main distributions. We binned the fixations in 20-ms bins based on their normalized latency. We calculated the proportion of fixations in each bin that landed on a face, creating a histogram of normalized saccadic latency against percentage of fixations landing on a face, the "saliency histogram" (Figure 3a).

We found that ESs are not attracted to faces above chance (Figure 3a; $p > 0.05$, Wilcoxon rank sum). We also found a significant increase in the percentage of fixations landing on the face starting 10 ms prior to the onset of the initiation of main distribution saccades ($p < 5.6 \times 10^{-43}$, Wilcoxon rank sum). This increased proportion of facial fixations is maintained throughout the main distribution, though it declines after 100 ms. The increased proportion of facial fixations starting at 10 ms prior to the onset of the main distribution is attributed to the fastest saccades of the main distribution being below the cutoff latency and to the discrete nature of the 20-ms bins. There are two facets to the main distribution: an early peak in facial saliency (20 ms in Figure 3a), followed by a general decline in facial saliency with increasing saccadic latency (40 to 200 ms in Figure 3a); 63.2% ± 1.3% (mean ± 95% confidence interval) of all MSs are to faces, highly above chance ($p < 10^{-15}$, Wilcoxon rank sum).

Fixations to text-containing images show a similar pattern (Figure 3b). None of the 19 observers made a single ES to text. Contrariwise, MSs are frequently made to text, with latencies as early as 120 ms. MSs also show high text saliency with 50.4% ± 4.2% of all fixations to text elements (mean ± 95% confidence interval). Thus,

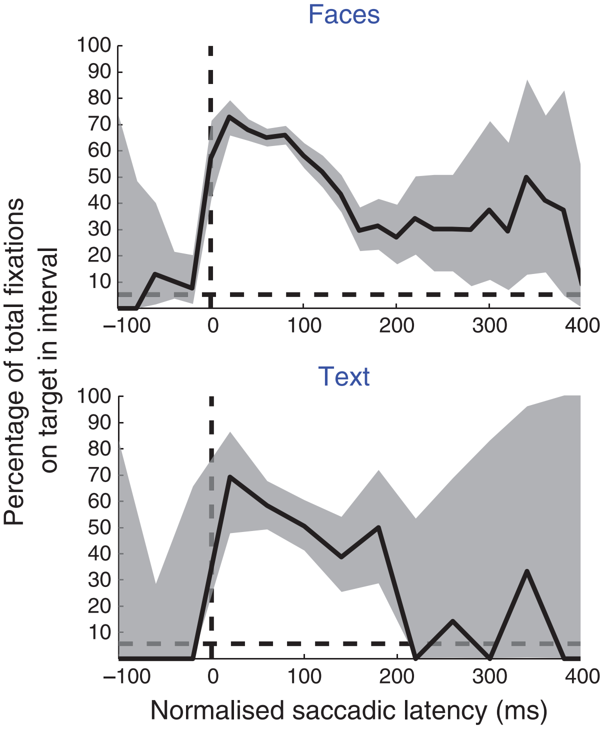

Figure 3. Saccades to specific image elements. Percentage of first saccades landing on (top) faces in 20-ms bins or (bottom) text in 40-ms bins during free viewing for all 19 subjects. Horizontal dashed lines represent chance performance (see Methods section). Vertical dashed lines mark the boundary of early (<0 ms) and main saccades (≥0 ms). Latencies are normalized to this breakpoint between the two distributions. The shaded area marks the 99% confidence intervals. ESs are not attracted to faces or text above chance, but main distribution saccades are highly selective.

faces and text fail to be salient to ES while attracting a large fraction of all fixations for MS (note that the images were chosen such that none contained both faces and text).

Inter-saccadic changes in decision mechanisms

While the first fixation is made to an unpredictable image, later fixations can take advantage of visual information gleamed during earlier fixations (transsaccadic integration of information; see Kotowicz et al., 2010 and discussion therein). We therefore set out to analyze the changes in saccadic decision from fixation to fixation (inter-fixation changes) for each individual subject. We pooled all the subject's saccades that landed on the face and broke those down by individual fixations (i.e., was it the first fixation they made, the second, third, etc.?). We considered these sets of saccades to faces as the output of a race-to-threshold model that decided whether or not to fixate onto a face during individual fixations. We used these to plot reciprobits for each fixation separately (Figure 3a).

The *gradient or steepness* is an inverse measure of prior probability for face viewing (Carpenter & Williams, 1995). High gradients correspond to low prior probability, or low expectation, of fixating a face (per unit time) and vice versa. Given our hypothesis about differences in image viewing between the first and subsequent fixations (unpredictable vs. predictable image), we can look at the gradients. These should be high for the first fixation and lower for subsequent ones.

We found that prior probabilities of the 2nd and subsequent fixations are higher than the first fixation. The gradient of the reciprobit in Figure 4a decreases after the first fixation. The gradient of the first fixation curve is greater by a factor of 1.7 relative to all subsequent fixation curves ($p < 10^{-13}$, 2-sample Kolmogorov–Smirnov). Across all subjects, the ratio of the first fixation to second fixation gradients is 2.3 ± 0.8 and the change is significant for all ($p < 10^{-3}$, 2-sample Kolmogorov–Smirnov).

This change in prior probability in later fixations is coupled to the emergence of early saccades that land on the face (Figure 4a). ESs fall on a line that reaches 50% at T = infinity, as in evoked saccade tasks, marking them out as a distinct population (Carpenter & Williams, 1995). There is an increase in the proportion of ES in second fixations (7.9% in face-containing images and 11.1% in text-containing images) as compared to the first fixation (2.9% in face-containing images and 3.4% in text-containing images; see Figure 3). This increased proportion of ES is also more selective for faces and text (77.9% of ESs are to faces and 66.2% to text) than those made to the first fixation (10.3% of ESs are to faces, while no ESs are to text; Figure 3). Thus, ESs evoked by image onset are not selective for faces or text, but ESs made to the second and subsequent fixations are highly selective.

As these changes in prior probability affect the median of the distribution, we considered the *intercept* with the T = infinity axis as a measure of the *mean rate of rise* of the decision signal of the face unit. The higher the *intercept,* or equivalently the more left-shifted the curve, the higher the *mean rate of rise* (Reddi & Carpenter, 2000). There is a change in the intercept, the mean rate of rise, from fixations one to two (Figure 4a). For the first fixation, the intercept is at 6.4 standard deviations from mean, dropping by the second fixation to 4.5 std, settling on 3.0 std by fixations three and four (2.6 std). From fixations two to three, we see a distinct rightward shift in the reciprobit, corresponding to a change in the mean rate of rise, independent of any change in prior probability ($p < 0.01$ for all subjects, 2-sample Kolmogorov–Smirnov). This shows that in the absence of any changes in the image or in the behavioral goal, dynamic changes occur in the mean rate of rise and therefore in the speed, and in the outcome, of saccadic decision.

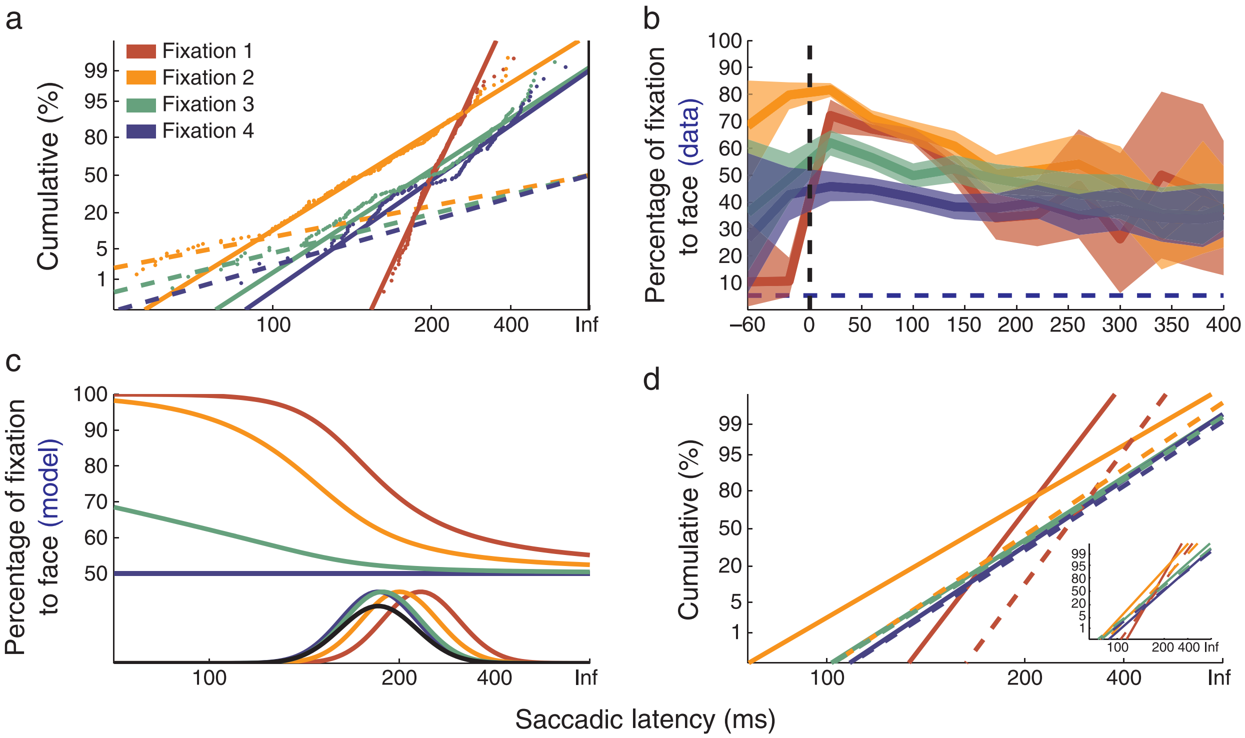

Figure 4. Variations in mean rate of rise between fixations and between objects determine the shape of saliency histograms through race-to-threshold competition. (a) Comparison of the first to the fourth fixation of a single subject (same as in Figure 2 and in following figures) of only those saccades that landed on a face. Solid lines represent trend lines of the main distributions; dashed lines are trend lines of the early distributions. (b) The proportion of fixations one through four on the face as a function of normalized saccadic latency. Latencies are aligned to the cutoff point between early and main distribution saccades as in Figure 2. Shaded areas represent 99% confidence intervals. At low latencies, the four distributions are significantly different but show similar values for longer latencies. (c) Modeling the effect of differences in the mean rate of rise between competing face and background decision units (bottom part of the panel; black distribution represents the face unit, while colored distributions represent background units of varying mean rate of rise) on the estimated proportions of fixations on the face (upper part of the panel). The distributions represent the normal probability density functions of the decision unit rising linearly to threshold at the given latency during any single trial, where the distribution mean represents its mean rate of rise. The smaller the difference in the mean rate of rise between distributions, the flatter the saliency histogram becomes. (d) Comparison of the trend lines of fixations of the main distributions to faces (solid lines) and to everything else (ground; dashed lines) for the 1st to 4th fixations for all subjects. Faces are seen as more salient and are viewed at a higher rate in faster saccades. The small box shows data for a single subject (same subject as top left).

Race-to-threshold competition

We hypothesized that differences in the mean rate of rise of decision units representing different objects within the visual field provides a means of biasing saccadic decision. The saliency histogram shows the proportion of saccades of a given latency that fall on a certain target and, thus, represents the probabilistic outcome of saccadic decision with respect to that target. Given the changes in the mean rate of rise between the *n*th and the (n + 1)th fixations (Figure 4a), we show that the saliency histogram for faces changes over the first four fixations, becoming progressively flatter as it loses the early peak in saliency. The attractiveness or saliency of faces for MS with shorter latencies (0–200 ms; Figure 4b), or short-latency saccades (SL-MSs), lessens and becomes equivalent to that for MS with longer latencies (>200 ms; Figure 4b), or long-latency saccades (LL-MSs). The difference between the saliency histograms for each fixation is significant (Figure 4b). This demonstrates a progressive change in the outcome of saccadic decision over the course of 4 saccades.

To understand how the changes in the mean rate of rise (Figure 4a) impact the shape of the curves in the saliency histogram (Figure 4b), we modeled race-to-threshold competition, as per the LATER model, mathematically. We used one face "unit" as a potential target of the race, competing with a ground "unit" referring to fixation to any possible alternative object, including the background, in the scene (Figure 4c). We confirm that when the face unit has a higher mean rate of rise than the ground unit, the proportion of fixations on the face is high for SL-MS and shows a gradual decrease toward LL-MS (Figure 4c). The larger the difference in the mean rate of rise between the face and the ground units, the steeper the change in proportion of fixations landing on the face from SL-MS to LL-MS is. Conversely, the closer the mean rates of rise of the two units, the flatter the saliency histogram.

Figure 4d shows reciprobit plots calculated with the data for both faces and ground. Interpreted in conjunction with Figure 4c, they explain the saliency histograms of Figure 4b. We show a large difference in the mean rate of rise between faces and the ground for the first fixation. This corresponds to a large proportion of SL-MS landing on faces with a steep drop in saliency toward LL-MS during the first fixation (Figure 4b). For the second fixation, the mean rate of rise of both the face and ground units has dropped, and the difference between the two has decreased. Thus, the change in the saliency histogram between SL-MS and LL-MS decreases. By the third and fourth fixations, there is almost no difference in saliency between faces and non-faces (ground), and thus, there is a correspondingly flat saliency histogram. The difference between the face trend lines and the ground trend lines is significant ($p < 0.05$, 2-sample Kolmogorov–Smirnov) for 14 of 19 subjects for the 1st fixation, 15 subjects for the 2nd fixation, 10 subjects for the 3rd fixation, and only 4 subjects for the 4th fixation.

Faces demonstrate an independent type of visual saliency

Changes in the mean rate of rise are attributed to the rate of incoming confirmatory sensory information (Reddi et al., 2003). However, we show these changes occurring in the absence of any changes in the image or information available to the subject. We hypothesized that the mean rate of rise was related to weighted, incoming sensory information and thus represented saliency. To investigate, we constructed a saliency map for each image using the Itti–Koch algorithm (Itti et al., 1998) with or without the addition of an equally weighted face conspicuity map that takes account of any faces in the image (Cerf et al., 2008).

Second fixations made to 600 images were separated on the basis of the saliency quartile in which they landed and plotted as reciprobits. Without an explicit face channel in the saliency algorithm, the curves representing fixations within the four saliency quartiles did not separate (Figure 5b; no significant differences in any subjects; $p > 0.05$, 2-sample Kolmogorov–Smirnov in all 19 subjects between all quartiles). However, when a standard face detection algorithm (Viola & Jones, 2001) was incorporated into the Itti–Koch saliency map, there was a significant separation of the curves (Figure 5a; $p = 3.2 \times 10^{-4}$ first to third quartiles, $p = 2.2 \times 10^{-5}$ first to fourth quartiles, 2-sample Kolmogorov–Smirnov). The most salient quartile produces a curve with the highest mean rate of rise. The mean rate of rise then reduces as the saliency is reduced. This separation in mean rate of rise is also present with later fixations. However, after the second fixation it occurs to a lesser degree, with little or no separation by the fifth fixation (separation between y intercepts of first and fourth quartiles, where the linear unit of the ordinate is standard deviations of a normalized Gaussian $N \sim [0,1]$; 1st fixation, 2.7 std; 2nd fixation, 0.7 std; 3rd fixation, 0.3 std; 4th fixation, 0.6 std; 5th fixation, 0.03 std) for the subject in Figure 4. This pattern is observed across all subjects (mean y intercept difference between first and fourth quartiles; 1st fixation, 4.5 std; 2nd

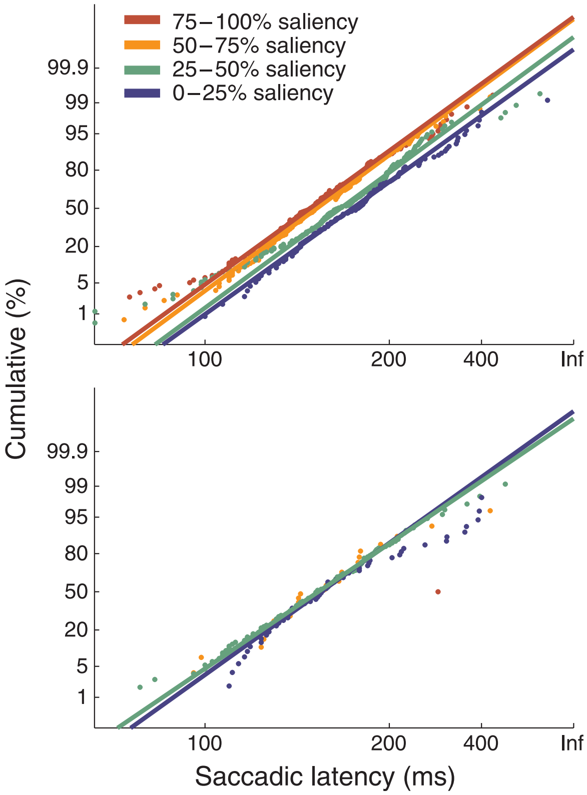

Figure 5. Inclusion of face detection in the standard Itti–Koch saliency algorithm correlates with separation of curves according to saliency. The saliency model with or without a face channel was applied to the 600 non-text-containing images viewed by subjects under the “free-viewing” condition. A single subject's fixations (same subject as in Figure 3) were divided into four groups according to the strength of saliency at the fixated location. Red points represent fixations to the most salient quartile of the saliency map, orange to the 50–75% quartile, green to the 25–50% quartile, and blue to the least salient quartile. (Top) The saliency map includes an explicit face detection channel (using the Viola–Jones face detection algorithm). The y intercept for T = infinity, and thus the mean rate of rise, correlates with saliency. The more salient the region, the higher the y intercept of the curve. (Bottom) Original Itti–Koch saliency algorithm without explicit face detection. When the saliency map is constructed solely from low-level visual features, the saliency level has little effect on the curve parameters.

fixation, 0.2 std; 3rd fixation, −0.04 std; 4th fixation, −0.2 std; 5th fixation, −0.2 std).

These results show that low-level visual features as typically used in saliency models: orientation, intensity, and color, do not drive significant changes in the mean rate of rise. It is the inclusion of a face detection pathway that leads to a correlation of saliency with mean rate of rise. In line with previous results, this demonstrates that it is specifically faces to which there is a significantly higher mean rate of rise, biasing saccadic decision and reducing latency, and not low-level visual features.

Top-down influences

In order to test top-down influences on saccadic decision, we analyzed a further experiment ("search task"), where the same subjects were directed to look at a face or other object prior to image onset and then had to search each image for this target. Since faces are salient, they end up looking at faces even when they were instructed to look for a non-face object. We quantified the proportion of fixations on the face, separated by the two directives given ("find an object: phone, Rubik's cube vs. "find a face"), and fixation number (Figure 6a).

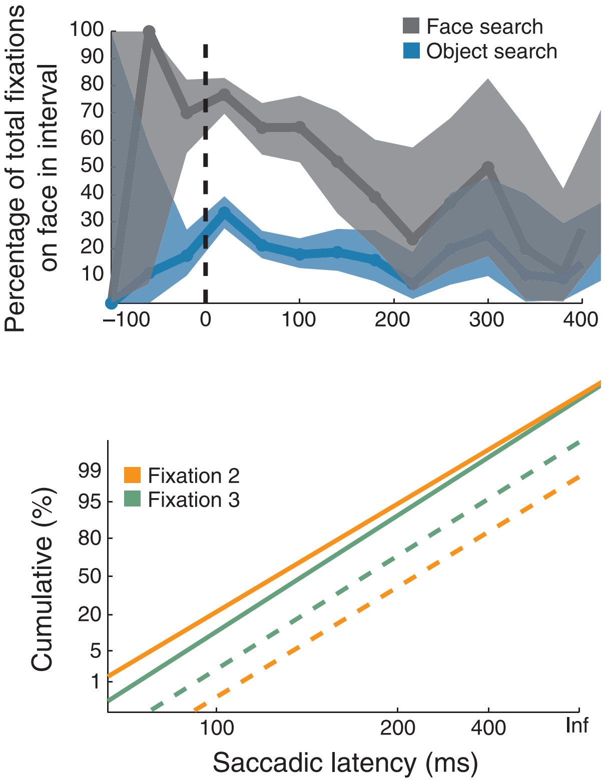

Figure 6. Top-down influences on facial fixations in the search task. (Top) Proportion of second fixations to faces when subjects were instructed to find a face (gray) or a non-face object (blue) in the image. Shaded areas represent 99% confidence intervals. (Bottom) Average slopes and intercepts for second (orange) and third (green) main fixations for 19 subjects (as these reflect fixations where the image was already seen by the subject beforehand) when instructed to fixate a face. The latency distributions of fixations that went to a face after it was already visited in an earlier fixation (dashed lines; goal previously completed) have lower mean rates of rise than the ones where the face was not visited earlier (solid lines; goal not previously completed). Therefore, the presence of a goal increases the mean rate of rise of units representing that goal in order to bias saccades toward the goal.

Faces are still visited significantly more than chance in both face and non-face object search tasks ($p < 0.01$, Wilcoxon rank sum; Figure 6a). Faces are fixated in 70.5% (59.1–80.3%, 95% confidence interval here and following) of early distribution saccades and in 62.7% (59.1–66.1%) of main distribution saccades when told to look for a face and 16.8% (9.7–26.0%) and 22.1% (19.7–24.8%), respectively, when told to look at an object. The differences between the two search tasks are highly significant for both early and main saccades ($p < 1 \times 10^{-16}$, $p < 1 \times 10^{-85}$, respectively, Wilcoxon rank sum).

However, due to the nature of the task, we cannot compare the results here to those in a free-viewing task as reciprobits. Instead, we looked at fixations to faces in a face search task (Figure 6b) under conditions of goal completion or non-completion. Fixating the face completed the goal of the task. However, subjects were then free to view the image for the remaining time. We therefore separated fixations made to faces during the subjects' second and third fixations on the basis of whether they had fixated the face in a prior fixation or not, i.e., whether or not they had already completed the task. We see a rightward shift in the curves upon goal completion, indicative of decreasing mean rate of rise. Specifically, the mean rate of rise of the face unit remains high until the face is fixated, at which point the goal of the task is completed. We therefore effectively have a free-viewing task, and we see a corresponding decrease in the mean rate of rise of the face unit (as demonstrated in Figure 4a).

Discussion

While the eyes are strongly and rapidly attracted to faces and text, one can imagine visiting those only once visual information is processed at a higher, cortical level where high spatial frequencies and emotional and semantic aspects of faces and text are represented (Oliva et al., 2003). This underlies the relatively longer latency of saccades to salient targets on image onset. Our results suggest a minimum latency of 120 ± 10 ms from the onset of a new image to a saccade to a face or to text (Figure 7), in agreement with previous studies in primates and human imaging (Kirchner & Thorpe, 2006; Liu, Harris, & Kanwisher, 2002; Mouchetant-Rostaing, Giard, Delpeuch, Echallier, & Pernier, 2000).

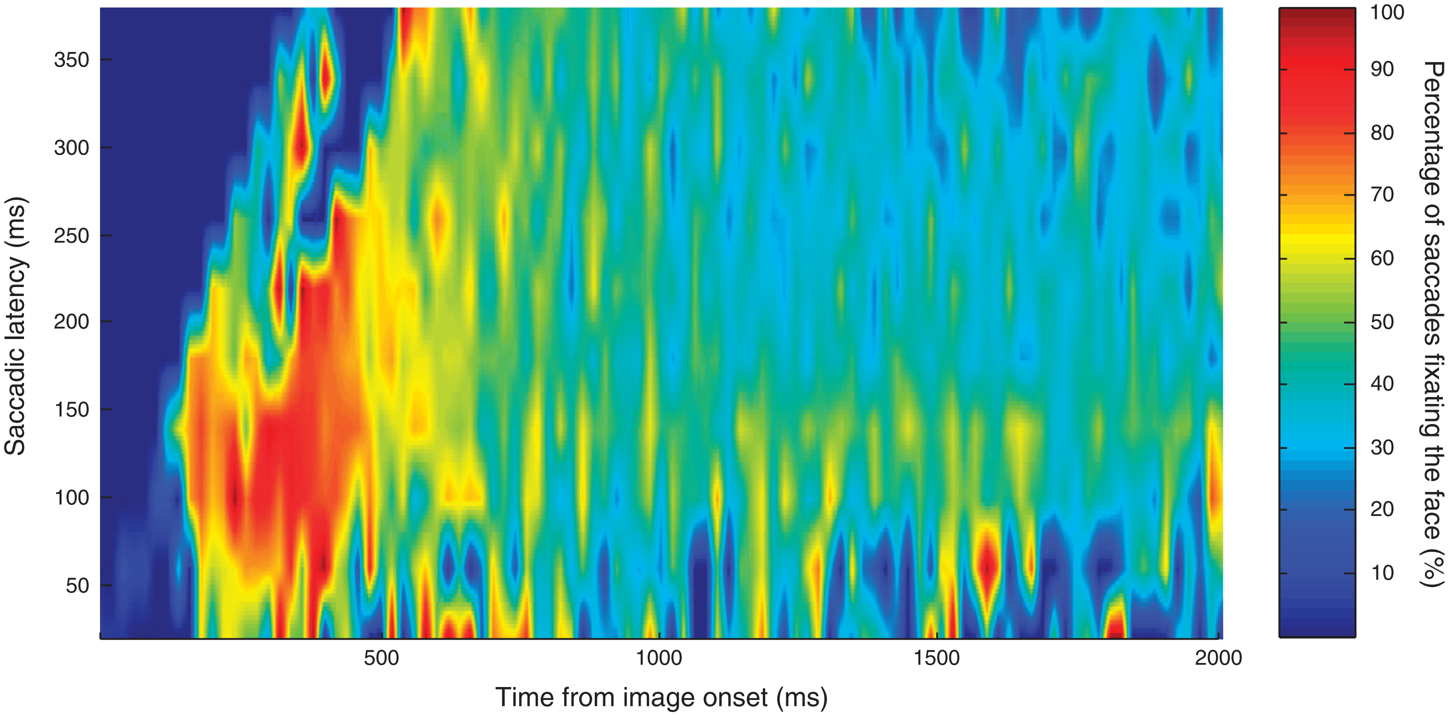

Figure 7. Proportion of fixations to face over the course of image viewing as a function of saccadic latency. Heat map showing proportion of fixations landing on faces as a function of saccadic latency and time after image onset. Warmer colors represent higher proportions. The *y*-axis represents a group of saccades of certain latency; the *x*-axis then represents the proportion of fixations to the face saccades of this latency make as a function of the time after image onset they were initiated at. Thus, the point (1000, 100) represents a saccade of latency 100 ms initiated 1000 ms after image onset. Saccades of 100–150 ms latency remain the most specific to faces throughout image viewing. Facial saliency is highest prior to 500 ms of image onset, and this applies to saccades of all latencies; 500 ms after image onset saccades with a latency of <100 ms and >200 ms become considerably less selective to faces. This demonstrates the absolute latency of facial detection at ~120 ms after image onset. Saccades of shorter latency only become attracted to faces after this time.

ESs are generated as early as 60–100 ms after the onset of the new retinal image. The short latency explains why this mechanism lacks access to object-specific "what" information if generated to image onset, as it takes visual information at least 90 ms to reach the higher areas of the ventral stream. Previous results from evoked saccade tasks suggest that ESs are not simply open-loop cortical saccades as they occur on first fixations to single targets. It was also shown that these saccades produced significantly more errors than typical saccades in target fixation (Reddi & Carpenter, 2000). Our results from ES evoked by image onset support these findings (Figure 7). ESs generated to image onset show no selectivity for faces or text.

In subsequent saccades to the same, static image, a distinct population of ESs emerges, capable of fixating salient objects strongly and rapidly. The emergence of these saccades is correlated with a significant change in image predictability as reflected in the reciprobit plots of later fixations, a factor known to influence the emergence of ES (Roos et al., 2008).

Based on the early latencies of saccades in free viewing of natural scenes, eye movements in rapid viewing are most likely the outcome of two different mechanisms of saccadic generation, one controlling ES and the other controlling the main distribution of saccades. Given the short latency of ES, the most likely candidates controlling their expression are subcortical circuits, in particular the superior colliculus (SC; Dorris, Olivier, & Munoz, 2007; Johnson, 2005).

We propose that these circuits provide a rapid, automatic mechanism for directing gaze to salient objects within a known visual scene without needing to wait for time-consuming cortical analysis of each new retinal image. Once a scene has been viewed, the SC could be "taught" where salient objects are by increasing the prior probability of the regions they are located in. This increases both the probability that an ES will be generated by outracing the main mechanism and also that it will target a salient object. This mechanism fits both with our results and with the known neurophysiology of the SC (Basso & Wurtz, 1998).

In terms of saccadic decision itself, we see significant changes in the mean rate of rise of units from fixation to fixation. There is an initial high, global mean rate of rise, which decays rapidly over a few fixations. As this occurs for all parts of a visual scene, this might best be interpreted as a novelty effect (Itti & Baldi, 2009; Ranganath & Rainer, 2003). The appearance of a novel stimulus would benefit from having an extremely high mean rate of rise to draw fixations as rapidly as possible. Such a bottom-up saliency signal should then die off quickly to free this mechanism up for more goal-directed, top-down influences (Fecteau, Bell, & Munoz, 2004).

Additionally, we see a difference in the mean rate of rise between face units and background, which also reduces over a few fixations. This correlates with, and explains, the changing shape of the saliency histogram over these fixations, representing changes in the outcome of saccadic decision. The mean rate of rise is not correlated with changes in low-level saliency but does correlate when a face detector is included in the saliency map. This suggests that the mean rate of rise of decision units for faces is higher by virtue of its semantic property and provides a means for biasing saccadic decision toward faces.

One can interpret these changes in the mean rate of rise of decision units, which rapidly decays over a second, as a signal coming from a bottom-up saliency map in the frontal eye fields (Moore & Armstrong, 2003) or the lateral intraparietal cortex (Bisley & Goldberg, 2003; Gottlieb, Kusunoki, & Goldberg, 1998; Kusunoki, Gottlieb, & Goldberg, 2000). Bottom-up saliency provides a means by which the sensory information encoding the most relevant objects is weighted. This increases the mean rate of rise of the relevant decision unit and generates a saccade to those objects quickest and most often. Mean rate of rise changes are also seen in top-down search tasks. We show that the mean rate of rise of the face in a face search task remains high until the face is fixated, at which point it reduces as in the free-viewing task. The presence of a goal then provides top-down influences for the same purpose (Thompson, Bichot, & Sato, 2005): By raising the mean rate of rise of a particular decision unit, the proportion of saccades targeted to the goal increases and the latency of these saccades decreases. In this sense, the mean rate of rise can be thought of as a *utility* signal (Good, 1952).

Both bottom-up saliency and the changes associated with goal completion in our task represent choices based on utility. The LATER model describes a mechanism for the inherent randomness in such decisions, which is arguably a survival advantage (Carpenter, 1999). We therefore propose that alteration of the mean rate of rise of such units, by bottom-up weighting of incoming sensory information and top-down influences, provides a common pathway for biasing random decision toward certain targets in a manner that reflects expected utility.

Acknowledgments

This work was funded by the Mathers Foundation, ONR, DARPA, and NSF.

Author contributions: Co-authors Moran Cerf and Michael Mackay contributed equally to this work.
Commercial relationships: none.

Corresponding author: Moran Cerf.
Email: moran@morancerf.com
Address: Caltech, 1200 East California Boulevard, Pasadena, CA 91125, USA.

References

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences, 103,* 449. [PubMed] [Article]

Basso, M., & Wurtz, R. (1998). Modulation of neuronal activity in superior colliculus by changes in target probability. *Journal of Neuroscience, 18,* 7519. [PubMed] [Article]

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115–147. [PubMed] [Article]

Bisley, J., & Goldberg, M. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. *Science, 299,* 81–86. [PubMed] [Article]

Carpenter, R. (1994). *Express optokinetic nystagmus*. Stuttgart, Germany: Georg Thieme.

Carpenter, R. (1999). A neural mechanism that randomises behaviour. *Journal of Consciousness Studies, 6,* 13–22.

Carpenter, R. (2001). Express saccades: Is bimodality a result of the order of stimulus presentation? *Vision Research, 41,* 1145–1151. [PubMed]

Carpenter, R., & Williams, M. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature, 377,* 59–62. [PubMed]

Cerf, M., Cleary, D. R., Peters, R. J., Einhäuser, W., & Koch, C. (2007). Observers are consistent when rating image conspicuity. *Vision Research, 47,* 3052–3060. [PubMed] [Article]

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12):10, 1–15, http://www.journalofvision.org/content/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article]

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems, 20,* 241–248. [Article]

Dorris, M., Olivier, E., & Munoz, D. (2007). Competitive integration of visual and preparatory signals in the superior colliculus during saccadic programming. *Journal of Neuroscience, 27,* 5053–5062. [PubMed] [Article]

Fecteau, J., Bell, A., & Munoz, D. (2004). Neural correlates of the automatic and goal-driven biases in orienting spatial attention. *Journal of Neurophysiology, 92,* 1728–1737. [PubMed] [Article]

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7*(1):10, 1–29, http://www.journalofvision.org/content/7/1/10, doi:10.1167/7.1.10. [PubMed] [Article]

Fischer, B., & Ramsperger, E., (1986). Human express saccades: Effects of randomization and daily practice. *Experimental Brain Research, 64,* 569–578. [PubMed] [Article]

Good, I. (1952). Rational decision. *Journal of the Royal Statistical Society (Methodological), 14,* 107–114. [Article]

Gottlieb, J., Kusunoki, M., & Goldberg, M. (1998). The representation of visual salience in monkey parietal cortex. *Nature, 391,* 481–484. [PubMed] [Article]

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49,* 1295–1306. [PubMed] [Article]

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 1254–1259.

Johnson, M. (2005). Subcortical face processing. *Nature Reviews Neuroscience, 6,* 766–774. [PubMed] [Article]

Khayat, P., Spekreijse, H., & Roelfsema, P. R. (2004). Correlates of transsaccadic integration in the primary visual cortex of the monkey. *Proceedings of the National Academy of Sciences of the United States of America, 101,* 12712. [PubMed] [Article]

Kirchner, H., & Thorpe, S. (2006). Ultra-rapid object detection with saccades eye movements: Visual processing speed revisited. *Vision Research, 46,* 1762–1776. [PubMed] [Article]

Kotowicz, A., Rutishauser, U., & Koch, C. (2010). Time course of target recognition in visual search. *Frontiers in Human Neuroscience, 4,* 12. [PubMed] [Article]

Kusunoki, M., Gottlieb, J., & Goldberg, M. (2000). The lateral intraparietal area as a salience map: The representation of abrupt onset, stimulus motion, and task relevance. *Vision Research, 40,* 1459–1468. [PubMed] [Article]

Leach, J., & Carpenter, R. (2001). Saccadic choice with asynchronous targets: Evidence for independent randomisation. *Vision Research, 41,* 3437–3445. [PubMed] [Article]

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: An MEG study. *Nature Neuroscience, 5,* 910–916. [PubMed] [Article]

Moore, T., & Armstrong, K. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature, 421,* 370–373. [PubMed] [Article]

Mouchetant-Rostaing, Y., Giard, M., Delpeuch, C., Echallier, J., & Pernier, J. (2000). Early signs of visual categorization for biological and non-biological stimuli in humans. *Neuroreport, 11,* 2521. [PubMed]

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. *Proceedings of the International Conference on Image Processing (ICIP), I,* 253–256. [Article]

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience, 4,* 193–202. [PubMed]

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural computation, 20,* 873–922. [PubMed]

Reddi, B., Asrress, N., & Carpenter, R. (2003). Accuracy, Information, and response time in a saccadic decision task. *Journal of Neurophysiology, 90,* 3538–3546. [PubMed] [Article]

Reddi, B., & Carpenter, R. (2000). The influence of urgency on decision time. *Nature Neuroscience, 3,* 827–830. [PubMed] [Article]

Roos, J., Calandrini, D. M., & Carpenter, R. H. S. (2008). A single mechanism for the timing of spontaneous and evoked saccades. *Experimental Brain Research, 187,* 283–293. [PubMed]

Schiller, P., Sandell, J. H., & Maunsell, J. H. (1987). The effect of frontal eye field and superior colliculus lesions on saccadic latencies in the rhesus monkey. *Journal of Neurophysiology, 57,* 1033. [PubMed] [Article]

Smith, P., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences, 27,* 161–168. [PubMed] [Article]

Thompson, K., Bichot, N., & Sato, T. (2005). Frontal eye field activity before visual search errors reveals the integration of bottom-up and top-down salience. *Journal of Neurophysiology, 93,* 337–351. [PubMed] [Article]

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113,* 766–786. [PubMed] [Article]

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 1,* 511–518. [Article]

von Neumann, J., & Morgensten, O. (1944). *Theory of games and economic behavior*. New Jersey: Princeton University Press.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum Press. [Article]