A confirmation bias due to approximate active inference

Ankani Chattoraj*[1], Sabyasachi Shivkumar*[1], Yong Soo Ra[1,2], Ralf M. Haefner[1]

[1]Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.
[2]Department of Bio and Brain Engineering, KAIST, Republic of Korea
Correspondence: achattor@ur.rochester.edu, sabyashiv@gmail.com, ralf.haefner@gmail.com.

Abstract

Collecting new information about the outside world is a key aspect of brain function. In the context of vision, we move our eyes multiple times per second to accumulate evidence about a scene. Prior studies have suggested that this process is goal-directed and close to optimal. Here, we show that this process of seeking new information suffers from a confirmation bias similar to what has been observed in a wide range of other contexts. We present data from a new gaze-contingent task that allows us to both estimate a participant's belief, and compare that to their subsequent eye-movements. We find that these eye-movements are biased in a confirmatory way. Finally, we show that these empirical results can be parsimoniously explained under the assumption that the brain performs approximate, not exact, inference, with computations being more approximate in decision-making compared to sensory areas.

Keywords: choice bias; perceptual decision-making; eye-movements; approximate inference

1 Introduction

Human decision-making is often biased, and few biases are as ubiquitous as the confirmation bias. While it has been documented across a wide range of cognitive and perceptual contexts, yet a unified understanding of its computational underpinning is currently missing (Nickerson, 1998; Michel and Peters, 2020). Two major components contribute to this bias: first, the biased *seeking* of information in supporting one's belief, and second, an *interpretation* of the observed information that is biased in the direction of one's existing beliefs. Over the past 15 years, several studies have documented evidence for a confirmation bias in perceptual decision-making tasks which have the benefit of allowing for the collection of large amounts of data using hundreds of repetitions from the same participants, and to study its neural basis. The evidence from those studies has shed light on the biased *interpretation* of sensory evidence (Nickerson, 1998; Michel and Peters, 2020; Lange et al., 2020). However, the *seeking* of new sensory evidence, most notably by eye-movements, has so far been found to be close to optimal (Najemnik and Geisler, 2005; Yang et al., 2016).

Our work makes two key contributions. First, it describes a new psychophysical discrimination task that requires observers to collect sensory information by making saccades. We designed the task to be able to measure how saccades may be influenced by existing beliefs. We found that the eye-movements of 8/10 participants were biased towards new information that agreed with the observers belief, reflecting a confirmation bias. Second, we show that this biased information-seeking behavior can be explained by a computational model that starts with an optimal Bayesian active sensing strategy (MacKay, 1992; Yang et al., 2016) but assumes that the required computations are implemented approximately via sampling. Importantly, such a model requires computing two terms – a sensory and a cognitive one – and it only displays the empirically observed confirmation bias when the number of samples used to compute the sensory term (and presumably implemented in sensory areas) is larger than the number of samples used to compute the cognitive term. Such a difference is compatible with previous observations on the dramatic difference in information capacity comparing sensory periphery and central processing, and suggestions that lower sensory areas act as a "high-resolution buffer" for higher-level computations (Lee and Mumford, 2003; Marois and Ivanoff, 2005).

*equal contribution

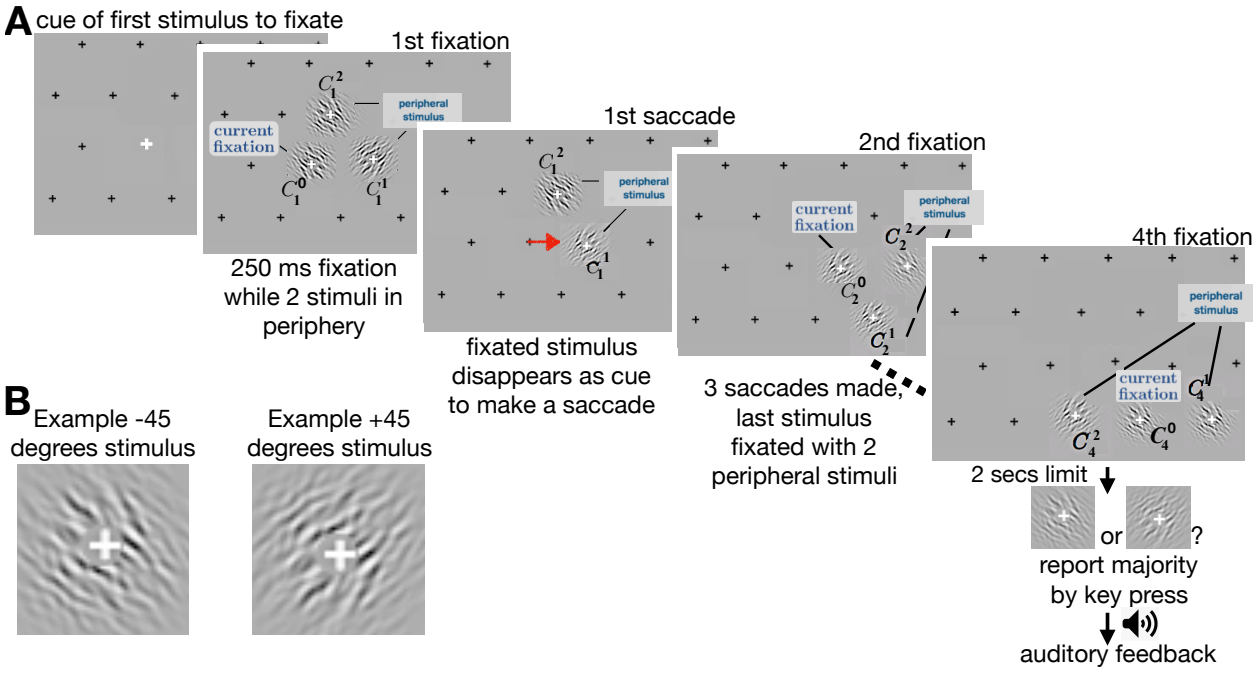

Figure 1: (A) One example trial of the gaze contingent task. A saccade is indicated by red arrow. The category on the fixated stimulus at any time t is denoted by C_t^0 and the categories of the peripheral stimuli are denoted by C_t^1 and C_t^2. If the first saccade is made to C_t^1, then it becomes C_{t+1}^0 for the next time step $t+1$. Note that the black crosses are used for illustration purposes to indicate possible locations of stimuli presentation. (B) Examples of stimulus used in the experiment.

2 Gaze contingent discrimination task

2.1 Rationale

Visual sensitivity to fine spatial structures differs greatly across the visual field. As a result, humans use saccades (as well as head and body movements) to move their eyes across a scene in order to collect information. In order to determine whether and, if so, how saccades are biased by previously collected information we designed a two-choice orientation discrimination task with a gaze-contingent stimulus display that allows for close control over the information present both at the current fixation point as well as in the periphery.

2.2 Task/Procedure

Participants were instructed to report the dominant (most frequent) orientation on the screen encountered while moving their eyes across a screen (Figure 1A). Each trial started with a fixation marker (white cross) in the center of a gray screen. After holding fixation for 200ms, three oriented stimuli appeared on the screen for a duration of 250ms (Figure 1B, details below): one stimulus around the fixation marker, plus two stimuli equidistant from fixation and each other. After 250ms, the stimulus at the fixation point disappeared providing a cue to the participant to make a saccade to one of the two stimuli in the periphery. While the saccade was in progress, the non-chosen peripheral stimulus disappeared and two new stimuli appeared, now peripheral with respect to the new fixation point. After 250ms, the fixated stimulus disappeared again and the participant had to make another saccade to one of the two peripheral stimuli. After a total of three saccades and four fixations, the participant reported their belief about the correct stimulus category for the entire trial. The orientation of each of the nine stimuli shown in each trial was drawn from the correct orientation category with probability 0.7. After each choice, auditory feedback was provided to the participant on whether their choice was correct. If a participant did not move their eyes within 200ms of the fixation stimulus disappearing, or if they did not report a choice within 2s after a trial ends, then it was aborted and ignored in analysis.

Participants were trained to perform the task using one block of 20 trials on the first day. Each following session consisted of blocks of 50 trials. Each of ten participants completed between 111 and 569 trials (median 250) across 3 sessions. The large variance in trial number resulted from the difficulty of the task, with a high fraction of aborted trials due to blinks, premature or delayed saccades, or saccades not landing at the center of one of the two peripheral stimuli. Aborted trials were excluded online, during the experiment.

2.3 Stimulus

Each stimulus was constructed by band-pass filtering Gaussian noise in the spatial frequency and orientation domains, and masking it by a soft-edged annulus (Beaudot and Mullen, 2006; Nienborg and Cumming, 2014; Bondy et al., 2018; Lange et al., 2020) (Figure 1B). Each annulus has a small white cross in the center which participants are instructed to foveate as shown in Fig 1B. Each stimulus subtends 2.08 degrees of visual angle around fixation. The centers of each peripheral stimuli lie at 2.88 degrees from each other and from the center of the fixated stimulus. The mean spatial frequency of the stimuli is $= 6.90$ cycles per degree, the spread of spatial frequency is $= 3.45$ cycles per degree, the (inverse) spread of orientation energy is 0.8, the image luminance $= 127 \pm 22$ and the width of the central annulus cutout is $= 0.43^{\circ}$. Stimuli were generated using Matlab and Psychtoolbox and presented on a gamma-corrected 1920x1080px 120 Hz monitor (Brainard, 1997). Participants kept a constant viewing distance of 105 cm using a chin-rest. Eye-movements were tracked using an Eyelink 1000.

Importantly, we chose the stimulus parameters and eccentricities in order to make the orientation of a foveated stimulus unambiguous for the participant, while providing some, but not perfect, information about the orientation of the stimuli in the periphery. If the information provided in the periphery is too low, then the brain will not be able to use it to decide where to move the eyes. If it is too high, then no new information is gleaned from moving the eyes and saccade plans may reflect different constraints than during natural viewing conditions. Furthermore, the design of each stimulus minimizes the effect of small fixational eye movements or variability in fixation location (within the annulus) on the information provided to the visual system.

2.4 Participants

The participants in this study consisted of 10 students at the University of Rochester (8 naive, 2 authors – highlighted in the analysis Figure 2A+B). Every naive participant was financially compensated for their time. All experiments were performed by following the guidelines and methods approved by the UR Research participants Review Board.

2.5 Analysis

Our task design allowed us to measure whether and how a participant combined their current belief about the correct task category with the information they expected in the periphery when determining where to move their eyes next.

Estimating a participant's current belief within a trials: We first performed logistic regression to determine the participant's choice bias as well as the weights assigned to the presented stimuli: 4 foveated stimuli and 5 non-foveated stimuli. This allowed us to estimate a participant's belief at the end of each fixation period, on each trial, by multiplying the stimuli presented so far with the corresponding weights and passing them through the logistic function yielding log odds(Figure 2C).

Estimate saccade bias: During each fixation period within a trial, the two peripheral stimuli were either of the same orientation (58%), or of different orientations (42%). In order to test for a confirmation bias in eye-movement strategy, we analyzed on the latter category – where saccades could be made either to a stimulus in agreement with our estimate of the participant's current belief, or disagreement (Figure 2B).

2.6 Findings

We found that participants could indeed successfully perform this challenging task and were consistent in their performance around threshold (Figure 2A). As expected, logistic weights on foveated stimuli (Figure 2C, black) are larger

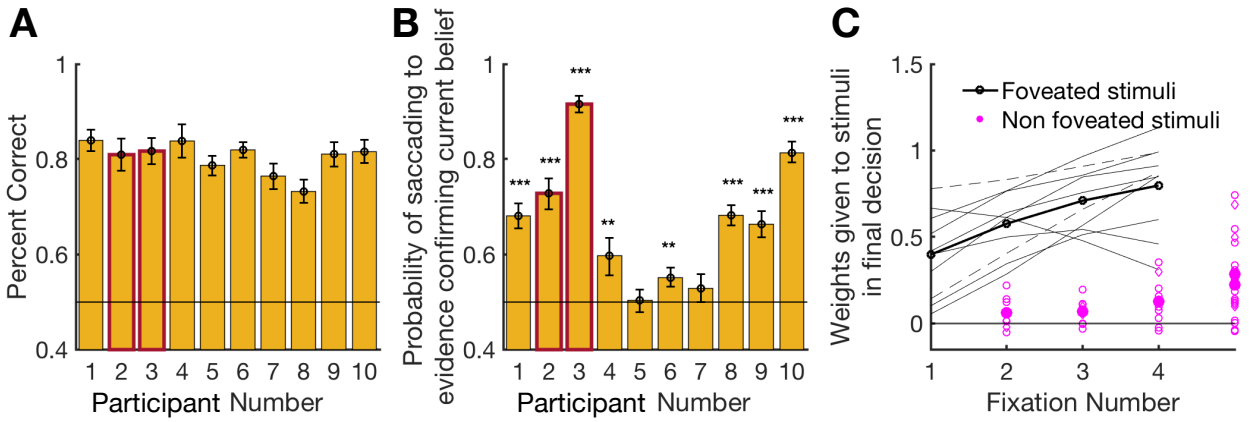

Figure 2: Participants show bias in eye-movements. (A) Performance of 10 participants in the experiment. (B) Probability with which participants saccade to confirmatory stimulus in the periphery based on already accumulated evidence (red bordered bars are for non naive participants) in both A and B. $p < 0.05$,$p < 0.01$ and $p < 0.001$ shown by 1, 2 and 3 stars respectively. (C) Weights given to stimuli in making the final choice. Thin black lines indicate individual participants (dashed lines for non naive). The thick black line indicates the mean weights across participants. Magenta hollow circles are for naive participants (diamonds for non-naive). The filled magenta circles indicate mean weights given to non-foveated stimuli across all participants.

than those on non-foveated ones (Figure 2C, magenta). This implies that for most participants the peripheral stimuli in our task contained some information that could in principle be used in planning saccades to stimuli expected to either confirm or dis-confirm one's existing belief. Furthermore, despite substantial participant-to-participant variability, most weights have a weakly increasing trend on average, in line with prior findings in comparable evidence accumulation tasks (Brunton et al., 2013; Wyart et al., 2012; Drugowitsch et al., 2016; Lange et al., 2020). This means that stimuli presented later in the trial have a slightly larger influence on average on the participant's choice than those presented earlier.

Importantly, we found that 9/10 participants were more likely to saccade to stimuli that agreed with their current belief about the trial category (8/10 statistically significant). One participant did not show any bias. It is possible that they could not extract enough information from the peripheral stimuli to guide their saccades, hence making saccades at random.

3 Approximate Bayesian active sensing model

3.1 Rationale

Maximal performance in our task is achieved by Bayesian active sensing, i.e. an observer who maximizes the gain in information about the correct choice with each saccade (MacKay, 1992; Najemnik and Geisler, 2005; Yang et al., 2016). However, it is straightforward to show that an *exact* Bayesian observer does not display any saccade selection bias since the gain in information is independent of stimulus orientation. However, what we will show below is that computing this gain in information *approximately*, in our case by sampling, will indeed induce an observer bias that matches our empirical data. The motivation for modeling the brain's approximate computations using sampling (as opposed to a variational approximation) is based on extensive prior work showing that sampling-based representations can account for a large amount of both cognitive (Griffiths et al., 2012; Gershman et al., 2012; Sanborn et al., 2010) and neural data (Fiser et al., 2010; Berkes et al., 2011; Haefner et al., 2016; Orbán et al., 2016; Echeveste et al., 2020). However, it is possible that a variational approximation entails the same qualitative bias as a sampling-based approximation (Lange et al., 2020).

3.2 Model details

Figure 3A shows a simplified version of the generative model for our task from the experimenter's perspective. Each trial is defined by a single category (45 degrees clockwise or counterclockwise), C, and consists of 4 sequential 250ms displays, indexed by $t = 1..4$, and represented by the plate (box) in Figure 3A. For the first display, $t = 1$, the orientations for all three presented stimuli are chosen independently from each other to agree with C with probability 0.7: one at the fovea, C_t^0, and two in the periphery, C_t^1 and C_t^2. The actually presented stimulus observed by the participant is then drawn as a Gaussian around the respective orientation modeling both the stochastic stimulus generation (orientation-filtered Gaussian noise) and the sensory noise of the visual system. The standard deviation of this Gaussian for the stimulus on the fovea is σ_{fovea}, and the variance in the periphery is $\sigma_{\text{periphery}}$, where $\sigma_{\text{fovea}} < \sigma_{\text{periphery}}$. For subsequent displays, only the two peripheral stimuli are drawn anew randomly to agree with C with 0.7 probability. The foveated stimulus, on the other hand, is identical to the peripheral stimulus saccaded to between the previous and the current display. This dependency between displays is not shown in Figure 3A for visual simplicity, but incorporated in our model. Importantly, optimal Bayesian inference over trial category C in this model requires optimally choosing saccade targets on each of the first three displays. As has been previously shown, this is accomplished by maximize the Bayesian Active Sensing (BAS) score across the two possible actions (MacKay, 1992; Najemnik and Geisler, 2005; Yang et al., 2016):

$$\max_{i=1,2} \quad \underbrace{\mathbb{H}\left[C_t^i|\mathcal{D}_t\right]}_{\text{sensory component}} \quad \underbrace{-\mathbb{E}_{p(C|\mathcal{D}_t)}\left[\mathbb{H}\left[C_t^i|C,\mathcal{D}_t\right]\right]}_{\text{cognitive component}} \tag{1}$$

where $\mathcal{D}_t = \left\{I_{1..t}^0, I_{1..t}^1, I_{1..t}^2\right\}$ represents all the stimuli presented so far. Intuitively, the first term represents the participant's uncertainty about the peripheral stimulus under consideration, and hence the information that could in principle be gleaned from saccading there. The 2nd component subtracts from that the information about this location that is already known given one's current belief about the trial category C. We call them 'sensory component' and 'cognitive component' since they are likely computed in a sensory and cognitive area, respectively.

In this work we hypothesize that the brain cannot compute this score exactly, but approximates it by sampling. We obtain:

$$\begin{aligned}
&\approx \quad \mathbb{H}\left[C_t^i|\mathcal{D}_t\right] - \frac{1}{n_{\text{cognitive}}} \sum_{l=1}^{n_{\text{cognitive}}} \mathbb{H}\left[C_t^i|C_{(l)},\mathcal{D}_t\right] \\
&\approx \quad \frac{1}{n_{\text{sensory}}} \sum_{c} \sum_{k=1}^{n_{\text{sensory}}} -p\left(C_t^i = c, C_{(k)}|\mathcal{D}_t\right) \times \log \frac{1}{n_{\text{sensory}}} \sum_{k'=1}^{n_{\text{sensory}}} p\left(C_{L1} = c, C_{(k')}|\mathcal{D}_t\right) \\
&- \quad \frac{1}{n_{\text{cognitive}}} \sum_{l=1}^{n_{\text{cognitive}}} \left[-\sum_{c} p\left(C_t^i = c|C_{(l)},\mathcal{D}_t\right) \times \log p\left(C_t^i = c|C_{(l)},\mathcal{D}_t\right)\right]
\end{aligned}$$

where n_{sensory} and $n_{\text{cognitive}}$ are the number of samples used to approximate each computation. In general, these two numbers can be different in the brain, reflecting differences in computational power and/speed in sensory and cognitive areas. For $n_{\text{sensory}} > n_{\text{cognitive}}$ the model exhibits a confirmation bias as described below.

3.3 Findings

We analyzed the behavior of our model and compared it to our empirical results. The three key parameters in our model are the numbers of samples used for both components, and the amount of sensory uncertainty in the periphery, $\sigma_{\text{periphery}}$. First, we found as expected that close-to-exact inference (large number of samples for both components) induced no bias in saccade choice regardless of any other parameters. Next, based on numerical simulations, we found the same result when $n_{\text{sensory}} = n_{\text{cognitive}}$. However, when $n_{\text{sensory}} > n_{\text{cognitive}}$ we found a bias for confirmatory saccades, while for $n_{\text{sensory}} < n_{\text{cognitive}}$ we found the opposite bias: to peripheral stimuli that disagreed with the participant's (model's) current belief. Since none of our participant showed the latter bias, we next focused on the case of $n_{\text{sensory}} \geq n_{\text{cognitive}}$ by fixing $n_{\text{sensory}} = 100$ (close to exact) while independently varying $n_{\text{cognitive}}$ and $\sigma_{\text{periphery}}$. The results on accuracy and saccade bias are shown in Figure 3. We found that for our discrimination task the performance depended almost exclusively on the sensory uncertainty, and only very weakly on the degree of the approximation used to compute the BAS score. On the other hand, as long as the sensory noise in the periphery was not too large,

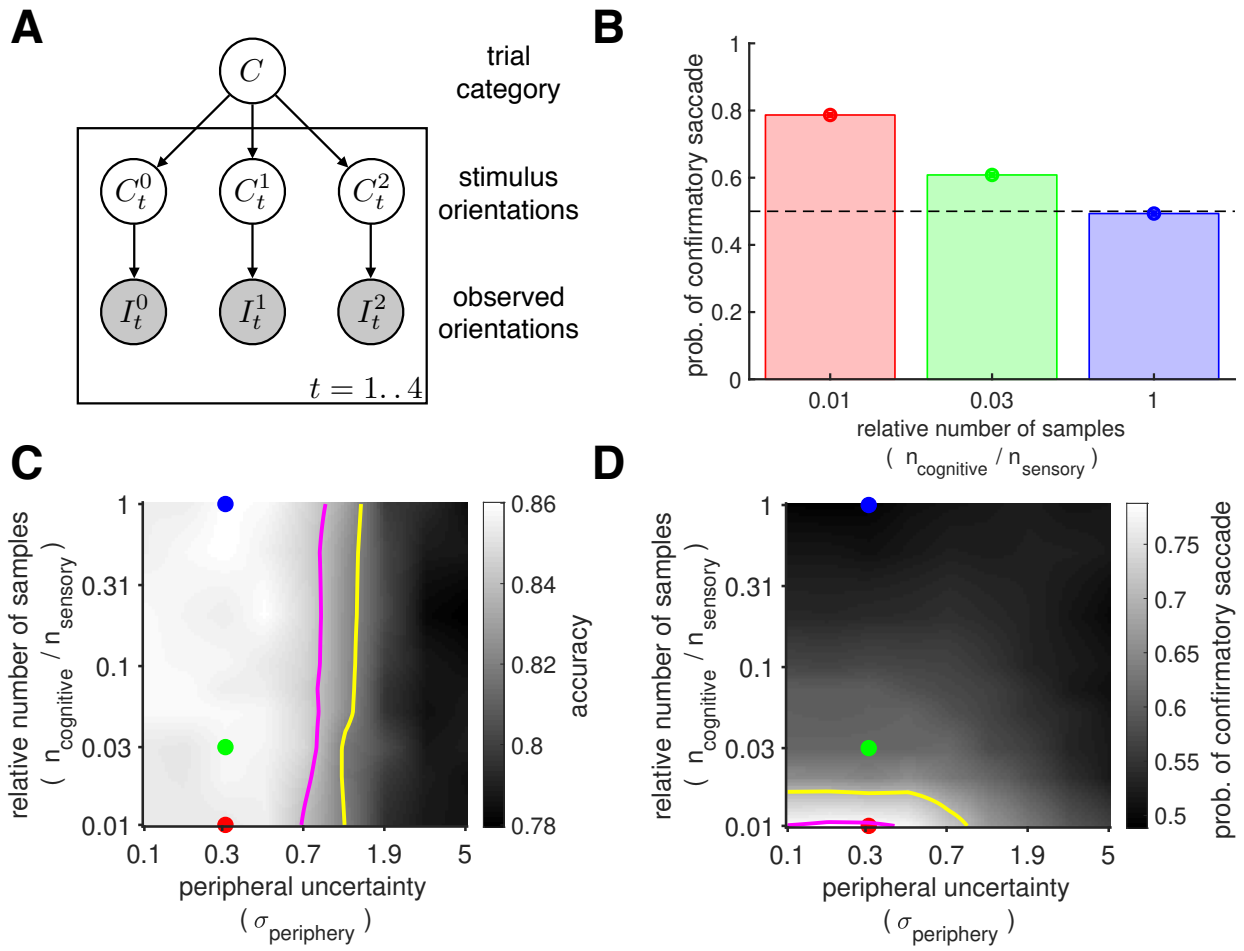

Figure 3: Approximate inference model of saccade selection (ref. model details for further information) (A) Simplified generative model of the task (B) Three simulated 'participants' who have a different degree of approximation parameterized by the relative number of samples in our model. The conformatory saccade selection biases observed in these 'participants' cover the range of empirically observed biases and is inversely related to the relative number of samples (C) & (D) Performance and conformatory saccade selection biases observed for different values of peripheral sensory uncertainty and relative number of samples. The parameters used for the three simulated 'participants' in (B) are shown in matching colors and the empirically observed performance and bias for two example subjects are shown as iso-contour solid lines

the strength of the bias depended primarily on the degree of approximation of the cognitive computation as quantified by $n_{\text{cognitive}}$: the coarser the approximation the larger the bias. Furthermore, the bias ranged from 0.5 to about 0.8 for the case $n_{\text{cognitive}} = 1$, covering the range of empirically observed values. The empirically observed performance and bias correspond to iso-contour lines (shown for two participants as solid lines) in the 2D parameter space shown in Figure 3C and 3D respectively. Three simulated 'participants' (models) are shown as red, green, and blue dots in Figure 3B. When the sensory noise in the periphery was too large, the bias disappeared due to the fact that regardless of approximation, the model could not infer any information to influence its saccades.

4 Discussion

Our work makes two main contributions. First, we provided evidence that our eye-movements are biased when sampling new information to make perceptual judgements. Second, we build on earlier work which showed that eye

movements are optimized to maximize information (Yang et al., 2016); and showed that empirically observed biases can be explained as the consequence of approximate computations in an ideal observer model.

We designed a psychophysics task where the participants had to make eye movements to collect information to determine the dominant (most frequent) orientation across all gratings. Each grating orientation matched the dominant orientation with a probability 0.7 and since every grating was drawn independently of each other, the ideal observer gathers information randomly to make its decision. We titrated the noise added to the stimulus such that the participant could not perceive the stimulus in their periphery reliably and therefore had to make eye movements to perform the task. While there was considerable across participant variability, most participants (8 out of 10) adopted a strategy to make eye movements that agreed with their accumulated belief. The accumulated belief was calculated by first inferring the weight of each stimuli to their final choice and then using this to infer intermediate beliefs. While such an analysis assumes that the weight assigned to each stimulus remains the same throughout the trial, we also looked at the the bias w.r.t. raw stimulus strength which also had the same qualitative pattern. We also note that a part of the variability across participants can be attributed to the variability in the stimuli sequence seen by the participants.

We modeled the observed biases as an ideal observer who chooses stimuli in the periphery that maximize the observer's information about the dominant orientation in the trial. This process is formalized as maximizing the mutual information between the peripheral stimulus orientation and the dominant trial orientation. The mutual information can be expressed as a difference between two components: (a) the "raw" peripheral stimulus information (the sensory component) and (b) the expected peripheral stimulus information based on the current belief about the overall dominant trial orientation (cognitive component). We hypothesize that participants approximate the underlying computations by sampling to compute each of the two components. A sampling based representation has been previously proposed for modeling behavior (Gershman et al., 2012; Griffiths et al., 2012) and neural responses (Fiser et al., 2010; Haefner et al., 2016; Orbán et al., 2016). We simulated this approximate information maximization model and found that a coarser approximation of the cognitive component than the sensory component led to a bias towards confirmatory saccades. On the other hand, we also found that a coarser approximation of the sensory than the cognitive component results in saccades to targets that are expected to disagree with the observer's current belief. Despite the model's ability to predict both types of biases in saccade selection we observed most participants having confirmatory saccades. We suggest that this is the result of the brain having a better representation of the sensory component. A finer representation in the early visual hierarchy has been previously proposed (Lee and Mumford, 2003). The bias in saccade selection also depends on the uncertainty associated with the peripheral stimulus orientation. If the peripheral information is very impoverished, then any degree of approximation mismatch does not lead to a bias as the observer cannot be biased by the peripheral information.

Eye movements are crucial for collecting information about the world. Our insights into how they are biased, and how approximate computations may be responsible for this bias, are not only important for our understanding of human vision but may also yield insights into potential biases and their causes in artificial intelligence systems.

Acknowledgments

This work was supported by NEI/NIH award R01 EY028811-01 (RMH, AC).

References

William HA Beaudot and Kathy T Mullen. Orientation discrimination in human vision: Psychophysics and modeling. *Vision research*, 46(1-2):26–46, 2006.

Pietro Berkes, Gergő Orbán, Máté Lengyel, and József Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, 2011.

Adrian G Bondy, Ralf M Haefner, and Bruce G Cumming. Feedback determines the structure of correlated variability in primary visual cortex. *Nature neuroscience*, 21(4):598–606, 2018.

David H Brainard. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997.

Bingni W Brunton, Matthew M Botvinick, and Carlos D Brody. Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340(6128):95–98, 2013.

Jan Drugowitsch, Valentin Wyart, Anne-Dominique Devauchelle, and Etienne Koechlin. Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, 92(6):1398–1411, 2016.

Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience*, 23(9):1138–1149, 2020.

József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–130, 2010.

Samuel J Gershman, Edward Vul, and Joshua B Tenenbaum. Multistability and perceptual inference. *Neural computation*, 24(1):1–24, 2012.

Thomas L Griffiths, Edward Vul, and Adam N Sanborn. Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21(4):263–268, 2012.

Ralf M Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016.

Richard D Lange, Ankani Chattoraj, Jeffrey M Beck, Jacob L Yates, and Ralf M Haefner. A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *bioRxiv*, page 440321, 2020.

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

René Marois and Jason Ivanoff. Capacity limits of information processing in the brain. *Trends in cognitive sciences*, 9(6):296–305, 2005.

Matthias Michel and Megan AK Peters. Confirmation bias without rhyme or reason. *Synthese*, pages 1–16, 2020.

Jiri Najemnik and Wilson S Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005.

Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

Hendrikje Nienborg and Bruce G Cumming. Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. *Journal of Neuroscience*, 34(10):3579–3585, 2014.

Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.

Adam N Sanborn, Thomas L Griffiths, and Daniel J Navarro. Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, 117(4):1144, 2010.

Valentin Wyart, Vincent De Gardelle, Jacqueline Scholl, and Christopher Summerfield. Rhythmic fluctuations in evidence accumulation during decision making in the human brain. *Neuron*, 76(4):847–858, 2012.

Scott Cheng-Hsin Yang, Mate Lengyel, and Daniel M Wolpert. Active sensing in the categorization of visual patterns. *Elife*, 5:e12215, 2016.