Selective disinhibition: A unified neural mechanism for predictive and post hoc attentional selection

CrossMark

Devarajan Sridharan [1,*], Eric I. Knudsen

Department of Neurobiology, Stanford University School of Medicine, Stanford, CA 94305, USA

ABSTRACT

The natural world presents us with a rich and ever-changing sensory landscape containing diverse stimuli that constantly compete for representation in the brain. When the brain selects a stimulus as the highest priority for attention, it differentially enhances the representation of the selected, "target" stimulus and suppresses the processing of other, distracting stimuli. A stimulus may be selected for attention while it is still present in the visual scene (predictive selection) or after it has vanished (*post hoc* selection). We present a biologically inspired computational model that accounts for the prioritized processing of information about targets that are selected for attention either predictively or *post hoc*. Central to the model is the neurobiological mechanism of "selective disinhibition" – the selective suppression of inhibition of the representation of the target stimulus. We demonstrate that this mechanism explains major neurophysiological hallmarks of selective attention, including multiplicative neural gain, increased inter-trial reliability (decreased variability), and reduced noise correlations. The same mechanism also reproduces key behavioral hallmarks associated with target-distracter interactions. Selective disinhibition exhibits several distinguishing and advantageous features over alternative mechanisms for implementing target selection, and is capable of explaining the effects of selective attention over a broad range of real-world conditions, involving both predictive and *post hoc* biasing of sensory competition and decisions.

1. Introduction

The world abounds with stimuli that are constantly competing for representation in the brain. Some of these stimuli are critical for guiding thought and behavior, whereas many others are irrelevant. Attention is the cognitive capacity that selects the most relevant information, at each moment in time, for prioritized processing and decision-making.

A variety of theories and computational models have been developed that describe the differential effects of selective attention on the prioritized processing of "target" versus other, irrelevant "distracter" stimuli. These models typically account for the effects of attention by invoking a bias in the competition among the competing stimulus representations to favor the neural representation of the target (Ardid, Wang, & Compte, 2007; Ardid, Wang, & Compte, 2007; Deco & Rolls, 2005; Desimone, 1998; Desimone & Duncan, 1995). These models of biased competition entail a particular sequence of processes: the selection signal, which biases competition, comes into play either before or while the target and distracter stimuli are present in the visual scene. This conventional scenario represents the prospective assignment of target priority or "predictive selection". However, what if a stimulus is identified as the target stimulus only *post hoc*, that is, after it has vanished from the visual scene? Conventional models cannot deal with this latter scenario.

Post hoc identification of targets for attention happens frequently in real-world situations. For example, transient, peripheral stimuli could be relevant for behavior, but these have often disappeared by the time attention is drawn to their location. Another common example is when a person is moving through the world and recognizes the importance of a stimulus only after she/he has moved away from it. Consider a researcher moving through a crowded room at a scientific meeting. Many faces appear briefly and disappear in her/his visual scene, and it may be a brief moment before a particular face is recognized as that of a former colleague. The researcher must retrospectively identify the location and heading

* Corresponding author at: 299 Campus Drive West, Rm D259, Fairchild Building, Department of Neurobiology, Stanford University School of Medicine, Stanford, CA 94305, USA.
E-mail addresses: dsridhar@stanford.edu, sridhar@cns.iisc.ernet.in (D. Sridharan).
[1] Present address: Centre for Neuroscience, Indian Institute of Science, Bangalore 560012, India.

of the colleague, while ignoring the movements of the other individuals in the room, in order to be able to track the colleague down. In both examples, information about the selected stimulus must be differentially processed *post hoc* (Carrasco, 2011; Liu, Pestilli, & Carrasco, 2005).

The effects of *post hoc* selection on information processing can be studied in the laboratory using psychophysical tasks that incorporate a *post hoc* response cue (Herrmann et al., 2010; Liu, Pestilli, & Carrasco, 2005; Pestilli et al., 2011). Such tasks are commonly referred to as "filtering" tasks (Palmer & Moore, 2009). In such tasks, competing stimuli are presented and have disappeared, before the identity of the relevant stimulus for decision-making is revealed to the subject by a cue (Fig. 1A–B). The subject is rewarded for making a correct decision about some feature of the *post hoc*-cued stimulus, such as its location, orientation, or direction of motion. By incorporating a predictive cue into this paradigm (e.g., Fig. 1A, red circle, or Fig. 1B, red arrow), the effects of both predictive cueing and *post hoc* cueing on stimulus discrimination can be studied (Pestilli et al., 2011).

Here, we develop a biologically inspired, dynamical model that, at its core, implements a recently reported neural mechanism: selective disinhibition (Fig. 1D). This mechanism controls local, feedforward inhibition of inhibition of sensory representations in the mammalian forebrain (Zhang et al., 2014). We show that this mechanism accurately accounts for attention's hallmark neurophysiological and behavioral effects, including target-distracter interactions, in both predictive and *post hoc* selection tasks (Palmer & Moore, 2009; Sridharan et al., 2014a; Zenon & Krauzlis, 2012). Our model presents a unified mechanism for attentional biasing of sensory representations and decision-making that can operate effectively in a wide range of real-world conditions.

2. Materials and methods

2.1. Task description

We developed the model to account for attention's effects in a filtering task, such as that shown in Fig. 1A. Following fixation, the subject is presented with a briefly flashed stimulus array consisting of two positive contrast dots, one on either side of the fixation point (stimulus encoding epoch). The elevation of each stimulus is independently sampled from values above or below the horizon. After a brief blank period (delay epoch), two response boxes appear to one side of the fixation point, one above and the other below the horizon. The subject must localize and indicate the elevation of the stimulus on the side of the response boxes by selecting the appropriate box, above or below the horizon (response epoch).

Two aspects of this target localization task must be emphasized. First, the response boxes constitute a *post hoc* cue, indicating which of the two stimuli was the target and which the distracter. Second, because the stimuli are no longer present when the identity of the target is revealed, the subject must retain the location of both stimuli to be able to successfully localize either one, *post hoc*. With this paradigm, the effects of predictive cueing on behavioral performance can also be measured, on interleaved trials, by introducing a predictive cue (Fig. 1A, red circle) that predicts the side of the target stimulus (the side of the response boxes) with high (~90–100%) validity.

The model is equally applicable to other types of filtering tasks, such as the one shown in Fig. 1B (Herrmann et al., 2010). In this task the stimuli are oriented Gabor gratings, and the subject must identify and report the orientation of the target grating as being clockwise or counter-clockwise of vertical in the presence of another distracter grating. In this task, the *post hoc* cue, indicating

the side of the target, is a spatial cue (box), whereas the predictive cue is a symbolic cue (Fig. 1B, red arrow).

2.2. Model equations

The filtering tasks shown in Fig. 1A–B consist of either 3 or 4 epochs: the (interleaved) predictive cueing epoch, the stimulus encoding epoch, the delay epoch, and the response epoch. We develop a single model that simulates each task epoch by switching flexibly between distinct dynamical modes.

Fig. 1C shows a schematic representation of the model. The model comprises of two neural populations (1 and 2) that mutually inhibit each other. One population (population 1) encodes the target stimulus; the other (population 2) encodes the distracter stimulus. Within each population, distinct sub-populations (A and B) encode the distinct values of each stimulus. In our "filtering" task (Fig. 1A), sub-population '1-A' is activated when the target stimulus is 'above' the horizon, and '1-B' is activated when the target stimulus is 'below' the horizon. Similarly, sub-populations '2-A' and '2-B' are activated when the distracter stimulus is located 'above' or 'below', respectively. Moreover, because each stimulus (target, distracter) appears above or below the horizon, but never at both locations, only one of the sub-populations within each population is activated on any given trial.

We model the evolution of target and distracter population activities on a given trial, with two mean-field equations, each representing the dynamics of one activated sub-population within each of the target and distracter populations, respectively:

$$dy_1/dt = -ky_1 - w_1^1 f(y_1) - w_1^2 f(y_2) + I_1 + I_{bkg}^1$$

$$dy_2/dt = -ky_2 - w_2^2 f(y_2) - w_2^1 f(y_1) + I_2 + I_{bkg}^2$$

where y_1 and y_2 are "state" variables that represent the net current flow into each neural population; the parameter k is a "leak" conductance that governs the rate of decay of each population's state (current); w_1^1 and w_2^2 represent the strength of inhibition (connection strength or synaptic weight) that each population receives from itself (recurrent inhibition); w_1^2 represents the strength of inhibition that the distracter population exerts on the target population (competitive inhibition), and vice versa for w_2^1; I_1 and I_2 represent the external input currents received by each population that represent, respectively, the strength of the target and distracter stimuli; I_{bkg}^1 and I_{bkg}^2 represent additive (possibly noisy) background current inputs to each population; $f(y)$ represents the "activation" function that describes the functional relationship between the input (current) and output (firing rate) of the population. $f(y)$ is modeled with a bounded, three-parameter sigmoidal hyperbolic-ratio (Naka-Rushton) function (Appendix A). Additional details regarding these variables and parameters of the model are provided in Appendix A.

These equations can be considered as "mean-field" approximations of the net currents into the two populations of neurons that mutually inhibit each other with current-based synapses (synapses that alter the currents into their postsynaptic neurons) (Bogacz et al., 2006; Bogacz et al., 2007). Modeling y_1 and y_2 as net population currents, rather than as neural firing rates as in previous models (Machens, Romo, & Brody, 2005; Wang, 2008), permits these state variables to assume both positive and negative values. The activation function, $f(y)$, that represents mean population firing rates, rectifies 'y', and assumes only non-negative values. We call $f(y)$ the "activity" of the respective target or distracter population.

The model simulates the various task epochs (predictive cueing, stimulus encoding, delay, and response) by systematic changes (local or global) to the leak conductance (k) and inhibitory connection strengths (w_i^j).