Active Inference and Learning in the Cerebellum

Karl Friston
k.friston@ucl.ac.uk
Wellcome Trust Centre for Neuroimaging, UCL, London WC1N 3BG, U.K.

Ivan Herreros
ivanherreros@gmail.com
Catalan Institute of Advanced Research, Technology Department,
Universitat Pompeu Fabra, Barcelona 08018, Spain

This letter offers a computational account of Pavlovian conditioning in the cerebellum based on active inference and predictive coding. Using eyeblink conditioning as a canonical paradigm, we formulate a minimal generative model that can account for spontaneous blinking, startle responses, and (delay or trace) conditioning. We then establish the face validity of the model using simulated responses to unconditioned and conditioned stimuli to reproduce the sorts of behavior that are observed empirically. The scheme's anatomical validity is then addressed by associating variables in the predictive coding scheme with nuclei and neuronal populations to match the (extrinsic and intrinsic) connectivity of the cerebellar (eyeblink conditioning) system. Finally, we try to establish predictive validity by reproducing selective failures of delay conditioning, trace conditioning, and extinction using (simulated and reversible) focal lesions. Although rather metaphorical, the ensuing scheme can account for a remarkable range of anatomical and neurophysiological aspects of cerebellar circuitry—and the specificity of lesion-deficit mappings that have been established experimentally. From a computational perspective, this work shows how conditioning or learning can be formulated in terms of minimizing variational free energy (or maximizing Bayesian model evidence) using exactly the same principles that underlie predictive coding in perception.

1 Introduction

Predictive coding is becoming a dominant paradigm in cognitive neuroscience, especially in understanding the computational architecture of perception. However, the same principles that underpin predictive coding can also be applied to action and choice behavior. This speaks to an embodied or an activist perspective on perception and the nature of self-organized behavior. There are several formal treatments of action and perception—for

example, distributed adaptive control and active inference (Friston, Mattout, & Kilner, 2011; Tani & Nolfi, 1999; Verschure, Voegtlin, & Douglas, 2003). In this work, we focus on active inference that provides a formal description of action and perception that can be implemented with predictive coding or Bayesian filtering (Adams, Shipp, & Friston, 2013; Friston et al., 2012). To highlight the embodied nature of active inference, we illustrate how it can furnish a detailed account of behavioral responses and, crucially, learning.

Our focus in this letter is on the simplest form of learning, Pavlovian learning, as exemplified by the eyeblink conditioning paradigm. Our objective is to show how the principles of predictive coding can be used to inform the computational anatomy of cerebellar circuits implicated in eyeblink conditioning. We also take the opportunity to show how predictive coding can account for learning through synaptic plasticity and show that predictive coding is not limited to perceptual inference in the cortex but can also be usefully applied to subcortical and cerebellar anatomy.

This letter comprises six more sections. Section 2 briefly reviews active inference and how it can be formulated as Bayesian filtering or predictive coding. This section establishes the central role of forward or generative models that enable inference about the causes of sensory inputs. Section 3 considers the minimal requirements of a generative model implied by eyeblink conditioning. In active inference, an unconditioned response (UR) corresponds to a prior belief that some latent or hidden state generates both action and an unconditioned stimulus (US). Pavlovian conditioning is the learning of the association between a conditioned stimulus (CS) and the unconditioned response. However, in eyeblink conditioning, we also have the constraint that the generative model must also account for spontaneous blinking. We model this using a generative model with itinerant dynamics (a heteroclinic cycle) that generates periodic blinks. This is an important aspect of the generative model because it introduces sequences or traces of activity that enable associations between conditioned stimuli and unconditioned responses that are separated in time (as in trace conditioning).

We then consider how this generative model might be implemented in the brain. This rests on the form of neuronal message passing implied by predictive coding, which specifies a particular pattern of connections. Using known anatomical (synaptic) connectivity, we then map the representations entailed by the generative model to various brain structures to maximize the consilience between empirical connectivity and the connectivity implied by predictive coding. This enables one to predict which connections are likely to mediate spontaneous blinking, startle responses, and (delay and trace) eyeblink conditioning. Section 4 uses simulations to illustrate these behavioral phenomena, thereby establishing the face validity predictive coding in this setting.

Section 5 addresses the predictive validity of the scheme using the known effects of selective brain lesions. These are simulated by deafferentation of

specific neuronal populations to emulate experimental manipulations (e.g., reversible deactivation). We find a remarkable correspondence between the lesion-deficit literature and the dissociation between delay and trace conditioning that emerges under predictive coding. For example, hippocampal lesions preclude trace conditioning but leave delay conditioning intact. This section also considers extinction and reproduces experimental attenuation of extinction with pharmacological lesions. We conclude with a brief summary of the implications of this formulation of Pavlovian conditioning.

2 Active Inference and Predictive Coding

There is a shift in cognitive science away from treating the brain as a passive filter of sensations, or stimulus-response link, toward a constructivist (and embodied) view in which the brain activity generates hypotheses or fantasies (*fantastic*: from the Greek *phantastikos*, the ability to create mental images, from *phantazesthai*) to be tested against sensory evidence (Gregory, 1968). This perspective dates back to the notion of unconscious inference (Helmholtz, 1866/1962) and has been formalized to cover deep or hierarchical Bayesian inference about the causes of our sensations and how these inferences induce beliefs, movement, and behavior (Clark, 2013; Dayan, Hinton, & Neal, 1995; Friston, Kilner, & Harrison, 2006; Hohwy, 2013; Lee & Mumford, 2003).

2.1 Predictive Coding and the Bayesian Brain. Modern formulations of the Bayesian brain, such as predictive coding, are among the most popular explanations for neuronal message passing (Clark, 2013; Friston, 2008; Rao & Ballard, 1999; Srinivasan, Laughlin, & Dubs, 1982). Predictive coding is a biologically plausible process theory for which there is a considerable amount of anatomical and physiological evidence. In these schemes, neuronal representations—in higher levels of cortical hierarchies—generate predictions of representations in lower levels (Friston, 2008; Mumford, 1992; Rao & Ballard, 1999). These top-down predictions are compared with representations at the lower level to form a prediction error (usually associated with the activity of superficial pyramidal cells). The resulting mismatch signal is passed back up the hierarchy to update higher representations (usually associated with the activity of deep pyramidal cells). This recursive exchange of signals suppresses prediction error at each level to provide a hierarchical explanation for sensory inputs. In computational terms, neuronal activity encodes beliefs or probability distributions over states in the world that cause sensations (e.g., my visual sensations are caused by a *face*). The simplest encoding corresponds to representing the belief with the expected value, or *expectation*, of a hidden or latent cause. These causes are referred to as *hidden* because they have to be inferred from sensory consequences. Figure 1 provides a formal and schematic illustration of predictive coding, which starts with a hierarchical or deep model of how sensations
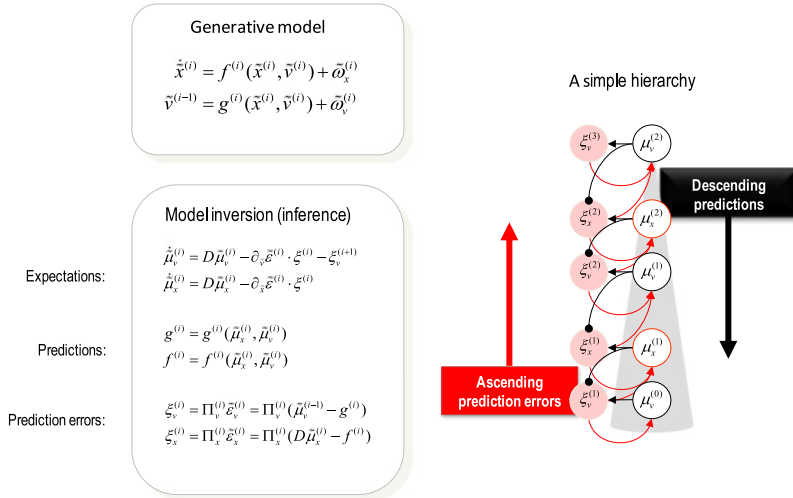
Generative model

$$\dot{\tilde{x}}^{(i)} = f^{(i)}(\tilde{x}^{(i)}, \tilde{v}^{(i)}) + \tilde{\omega}_x^{(i)}$$
$$\tilde{v}^{(i-1)} = g^{(i)}(\tilde{x}^{(i)}, \tilde{v}^{(i)}) + \tilde{\omega}_v^{(i)}$$

A simple hierarchy

Descending predictions

Model inversion (inference)

Expectations:
$$\dot{\tilde{\mu}}_v^{(i)} = D\tilde{\mu}_v^{(i)} - \partial_{\tilde{v}}\tilde{\varepsilon}^{(i)} \cdot \xi^{(i)} - \xi_v^{(i+1)}$$
$$\dot{\tilde{\mu}}_x^{(i)} = D\tilde{\mu}_x^{(i)} - \partial_{\tilde{x}}\tilde{\varepsilon}^{(i)} \cdot \xi^{(i)}$$

Predictions:
$$g^{(i)} = g^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)})$$
$$f^{(i)} = f^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)})$$

Prediction errors:
$$\xi_v^{(i)} = \Pi_v^{(i)}\tilde{\varepsilon}_v^{(i)} = \Pi_v^{(i)}(\tilde{\mu}_v^{(i-1)} - g^{(i)})$$
$$\xi_x^{(i)} = \Pi_x^{(i)}\tilde{\varepsilon}_x^{(i)} = \Pi_x^{(i)}(D\tilde{\mu}_x^{(i)} - f^{(i)})$$

Ascending prediction errors

Figure 1: Hierarchical neuronal message passing in predictive coding. In these schemes, neuronal activity encodes expectations about the causes of sensory input, such that expectations minimize prediction error. Prediction error is the difference between (ascending) sensory input and (descending) predictions of that input. Minimizing prediction error rests on recurrent neuronal interactions among different levels of the cortical hierarchy. (Left) These equations represent the neuronal dynamics implicit in predictive coding. The equations in the upper panel describe a generative model in terms of equations of motion that describe dynamics of hidden states in the world. This motion is influenced by hidden causes that are themselves generated by hidden states at a higher hierarchical level. When this model is used to account for sensory data, the (Bayes) optimal expectations of hidden states and causes can be written as the solution to a gradient descent on variational free energy. When this gradient descent is expressed in terms of prediction errors, we get the equations in the lower panel. The form of these equations implies the message passing scheme shown on the right. (Right) Prediction errors at the ith level of the hierarchy are simply the difference between expectations encoded at that level and top-down predictions of those expectations. The expectations per se (white circles) are driven by prediction errors (pink circles) so that they reduce the sum of squared (precision weighted) prediction error. Ascending prediction error signals are shown as red connections, while descending predictions are shown in black. Because prediction error computation involves subtracting predictions from expectations at lower levels, the descending black connections are inhibitory. See the appendix for a detailed explanation of these equations and the variables in this figure.

are generated (known as a generative model). Given a generative model, one can then write down the equations of motion for the expectations that maximize Bayesian model evidence (or minimize prediction error). This is known as Bayesian filtering or predictive coding.

2.2 Predictive Coding, Behavior, and Learning. This formulation of neuronal processing accounts for perception; however, the underlying imperative to minimize prediction error can also be applied to action. This follows from the fact that all behavior is mediated by classical (or autonomic) reflexes—that can be cast as minimizing proprioceptive (or interoceptive) prediction errors. In other words, by simply equipping predictive coding with reflexes, we have a simple description of embodied or active inference. In this view, motor commands are replaced by descending proprioceptive predictions that elicit movement by engaging reflexes (that are preprogrammed to minimize proprioceptive prediction error). We will use this later to model unconditioned responses that are generated by motor commands whose corollary discharge coincidentally predicts unconditioned stimuli. This provides a slightly unusual view of reflexive behavior: stimuli do not cause responses but induce expectations that generate predictions in both the motor (proprioceptive) and sensory (interoceptive or exteroceptive) domains. The former produce movement, while the latter suppress sensory prediction errors. In both cases, prediction error is minimized.

The suppression of prediction error also explains learning and plasticity. In predictive coding, hierarchical expectations (and their associated prediction errors) are generally thought to be encoded by synaptic activity. However, there are parameters of the generative model that change slowly over time. Expectations about these parameters are thought to be encoded by synaptic connectivity. This means that synaptic connections should change slowly to minimize prediction error (accumulated over time). When expressed mathematically, these changes have the form of associative or Hebbian plasticity, in which the connection strengths are driven by (postsynaptic) prediction error responses (see the appendix).

In what follows, we use all three processes to minimize prediction error (and implicitly maximise Bayesian model evidence): perception, action, and learning. A more detailed and technical description of these processes is provided in Figure 1 and the appendix. The key thing to remember is that having specified a generative model, the equations underlying perception, action, and learning are specified uniquely and automatically. These equations of motion can then be integrated to simulate the neuronal and behavioral responses that mediate the Bayesian belief updating implicit in predictive coding. In the next section, we focus on a generative model of eyeblinks and their conditioning on various stimuli.

3 The Generative Model

This section describes the generative model used in subsequent sections to understand the functional anatomy of spontaneous blinking and eyeblink conditioning—and to simulate the effect of various lesions. In brief, we tried to identify a minimal model that could accommodate spontaneous blinks, reflexive eyeblinks to unconditioned stimuli, and the learning of temporal associations between conditioned and unconditioned stimuli. The

particular model we describe is not necessarily unique; however, it is suffi-
cient for our purposes.

The model has two hierarchical levels, with the hidden states at the first
level modeling beliefs about conditioned stimuli, eyeblinks, and uncondi-
tioned stimuli. At this level, proprioceptive predictions (motor commands)
are generated by expectations about spontaneous eyeblinks and uncondi-
tioned stimuli. Hidden states at the second level embody a sense of time
through Lotka-Volterra dynamics. These dynamics can be regarded as a
central pattern generator that visits a succession of unstable fixed points
in a particular sequence. Technically, this is known as a heteroclinic cy-
cle, which generates a sequence of latent or hidden states. These states are
passed to the first level via a softmax function and generate spontaneous
eyeblinks or predict conditioned stimuli. In this example, the heteroclinic
cycle has an inset or sequence of (three) states that lead to the cycle. The
cycle per se generates spontaneous eyeblinks periodically, while the first
state of the inset generates a conditioned stimulus. The subsequent unsta-
ble fixed points play the role of trace for echo states that enable the learning
or association of subsequent unconditioned stimuli (US) that follow the
conditioned stimuli (CS) some time later.

Learning corresponds to Hebbian plasticity (that minimizes prediction
error) in the connections between expectations about a US and (first level)
expectations about the CS for delay conditioning and (second level) echo
states for trace conditioning. These are shown as broken lines in Figure 2.
Figure 2 also provides the equations of motion and mappings from hidden
causes to hidden states that define this hierarchical generative model. This
figure uses the formalism of Figure 1, which is described in detail in the
appendix.

4 The Functional Anatomy of Eyeblink Conditioning

Having established the basic (minimal) form of our generative model, we
now have a computational architecture for the corresponding Bayesian
filtering or predictive coding scheme. This architecture is shown in terms of
posterior expectation and prediction error units (associated with neuronal
populations) and connections that mediate the influence of expectations on
error units and vice versa (associated with connections among populations).
Notice that this connectivity is rather sparse and conforms to some simple
rules (see the appendix): (1) every expectation unit is reciprocally connected
to a companion error unit, (2) expectation units send connections only to
prediction error populations and vice versa, (3) ascending connections are
exclusively prediction errors that are passed forward to inform expectations
at the higher level, (4) descending connections are exclusively predictions
that are subtracted from expectations in lower levels and are therefore
designated as inhibitory, and (5) reciprocal connections are restricted to a
hierarchical level or link neighboring levels. With these rules in mind, can
we now use the known connectivity of eyeblink conditioning circuits to
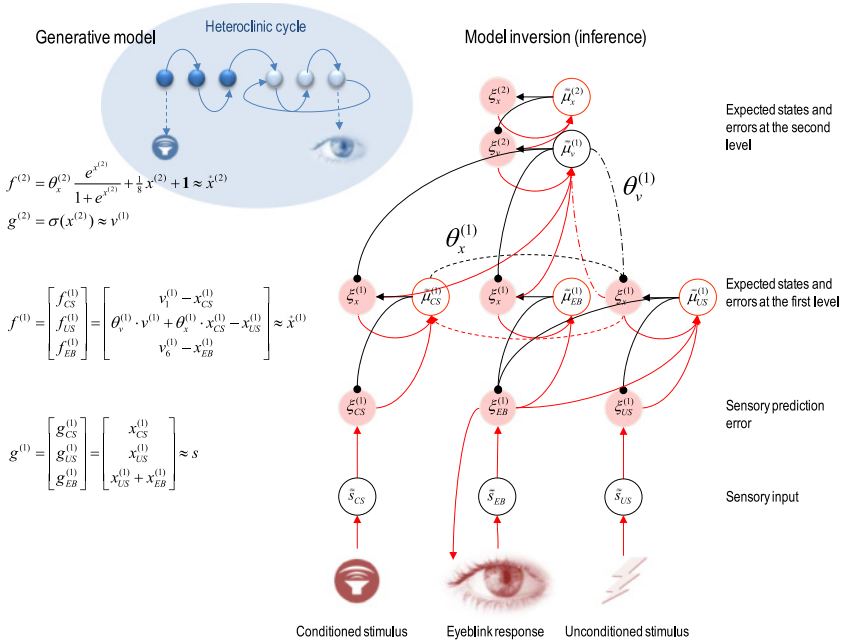
assign these expectations and error units to specific anatomical nuclei and neuronal populations?

Figure 3 provides a schematic of the key connections involved in eyeblink conditioning (lower right panel). In brief, the inferior olive (IO) receives excitatory sensory afferents from the trigeminal nucleus reporting somatosensory or nociceptive stimuli (e.g., an unconditioned stimulus like an air puff or electrical shock). The inferior olive then sends ascending excitatory (climbing fiber) connections to the Purkinje cells (PC) of the cerebellar cortex (CC) and interpositus nucleus (IP). The Purkinje cells reciprocate inhibitory connections to the inferior olive through the interpositus nucleus. The interpositus nucleus provides descending motor control signals via the red nucleus (RN) and facial nucleus to elicit an eyeblink. In addition, the interpositus nucleus receives excitatory afferents from the pontine nuclei (PN), which themselves send ascending excitatory (mossy fiber) connections to granular (and Golgi) cells (GC) in the cerebellar cortex. These provide excitatory inputs to Purkinje cells (via parallel fibers). Eyeblink conditioning is thought to involve plasticity (long-term depression) in connections to and from the cerebellar cortex, particularly the output nucleus of the cerebellum: the interpositus nucleus (Freeman & Steinmetz, 2011; McCormick & Thompson, 1984; Yang, Lei, Feng, & Sui, 2014). Spontaneous blinking is thought to be generated by (globus pallidus) eyeblink centers that project to the interpositus nucleus via the pontine nuclei. It is

also known that the hippocampus is involved in (conditioned) eyeblinks (Kim, Clark, & Thompson, 1995). For example, lesions to the hippocampus abolish trace conditioning, something that we will return to later.

The left panel of Figure 3 shows the same predictive coding architecture described in Figure 2 but rearranged to match the predictive coding connections to the anatomical connectivity summarized above. There are possibly many rearrangements—and implicit anatomical assignments—we could have considered. The designations in Figure 3 associate exteroceptive, proprioceptive, and somatosensory prediction errors with the pontine, red, and inferior olivary nuclei, respectively. Expectations of hidden states (and their prediction errors) mediating spontaneous eye blinks, unconditioned stimuli, and conditioned stimuli have been assigned to the

---

Figure 2: This figure describes the particular generative model of eyeblink conditioning used in subsequent simulations. The highest level of the generative model entails (six) hidden states that show itinerant dynamics, prescribed by Lotka-Volterra equations of motion. A softmax function of these hidden states generates hidden causes that produce fluctuations in (three) hidden states of a lower sensory level. These hidden states correspond to latent variables mediating conditioned stimuli (CS), eyeblink responses (EB), and unconditioned stimuli (US). The hidden states generate hidden causes mediating (predicted) sensations; where the CS is in the exteroceptive domain (e.g., a loud tone), the US generates interoceptive or somatosensory signals (e.g., a nociceptive stimulus) and proprioception associated with an eyeblink. Finally, the eyeblink response predicts the proprioceptive consequences of an eyeblink. The equations on the left prescribe the precise form of these hierarchical contingencies. The first hidden state at the second level of the hierarchy generates a conditioned stimulus, while the last (sixth) state generates a spontaneous eyeblink. Crucially, the itinerant (heteroclinic) dynamics have been chosen to produce a heteroclinic cycle among the last three states. This means, in the absence of any conditioned stimuli, that the animal will cycle through the last three states, emitting a spontaneous eyeblink on every cycle. However, if there is a CS, predictive coding will activate the first state. This state is an unstable fixed point and will yield to the second and subsequent states until the heteroclinic cycle is reentered. Effectively, this resets the heteroclinic cycle or central pattern generator, providing a sequence of (echo) states following a CS. This heteroclinic cycle is shown schematically in the upper inset. Note that the dynamics of the hidden state mediating a conditioned response can be affected by the hidden states generating a CS or the hidden causes associated with the (echo) states at the higher level. These influences depend on (connectivity) parameters that have to be learned, which are depicted as broken lines above. The right panel shows the predictive coding scheme implied by the equations of the generative model in the left panel. This schematic follows the same format of the previous figure, with (ascending) prediction error connections in red and (descending) predictions in black.
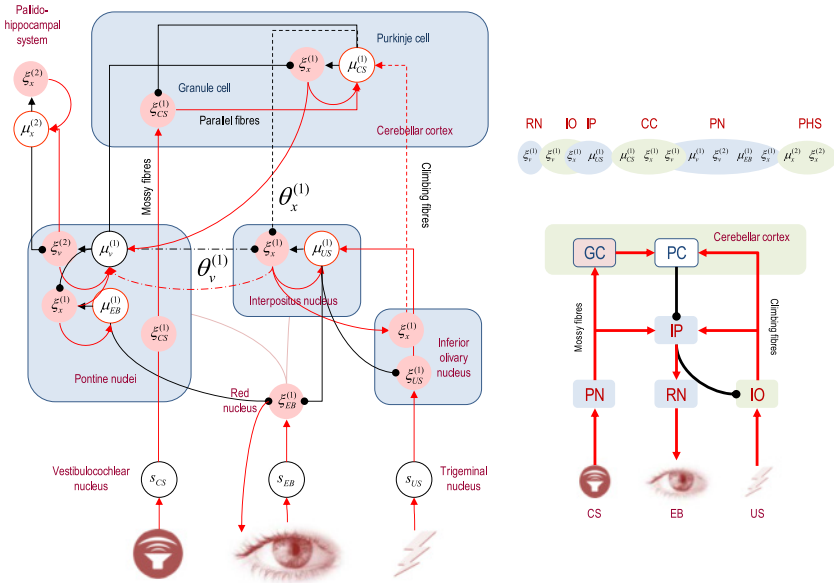
pontine nuclei, interpositus nucleus, and cerebellar cortex, respectively. Finally, the hidden states at the second level have been associated with the globus pallidus and the hippocampus (designated by PHS) or amygdala. The associated (second level) hidden causes have been placed in the pontine nuclei.

Note that we have duplicated two populations to conform to the anatomical connectivity constraints: excitatory principal cells encoding prediction errors in the interpositus nucleus have been replicated in the inferior olive. Similarly, exteroceptive prediction error units are replicated in the pontine nuclei and cerebellar cortex (as granule cells). With these duplications, there is a remarkable correspondence between the empirical and predictive coding connectivity. Of special note are the inhibitory (descending) connections from the cerebellar cortex to the interpositus nucleus and from the interpositus nucleus to the inferior olive. Note that the connections implicated in learning or conditioning (broken lines) implicate (recurrent) connections between Purkinje cells and the interpositus nucleus and (recurrent) connections between the pontine nuclei and interpositus nucleus. This is consistent with the key role of the interpositus nucleus in eyeblink conditioning.

This putative mapping between the predictive coding and neuroanatomy suggests the functional anatomy depicted in Figure 4: spontaneous blinking (upper left) mediated by descending predictions (from the eyeblink center in the globus pallidus) to the pontine and red nuclei. Conversely, when blinking is elicited reflexively by an unconditioned stimulus,

the descending proprioceptive predictions (or motor commands) emanate from the interpositus nucleus. In active inference, a reflex of this sort rests on expectations about the joint occurrence of an unconditioned stimulus and response. This expectation is induced by the unconditioned stimulus and subsequently elicits action through descending predictions of expected proprioceptive consequences. It can be seen (upper right) that these descending predictions target the red nucleus and inferior olive, providing motor commands and corollary discharge, respectively.

After learning or conditioning, we can now call on further connections to mediate conditioned responses. However, the form of the conditioning implicates different connections. In delay conditioning, the unconditioned and conditioned stimuli are coextensive in time, where the conditioned stimulus usually starts before the unconditioned stimulus. It is therefore sufficient to expect contemporary exteroceptive and proprioceptive eyeblink signals. This calls on expectations of conditioned stimuli encoded by

---

Figure 3: This figure reproduces the predictive coding scheme of Figure 2 but assigns the expectation and error populations to particular anatomical structures. This neuroanatomical designation is based on the connectivity shown on the lower right. (Lower right) This schematic summarizes the canonical connectivity of eyeblink conditioning. The inferior olive (IO) receives excitatory sensory afferents from the trigeminal nucleus reporting somatosensory or nociceptive stimuli (e.g., an unconditioned stimulus). The IO then sends ascending excitatory (climbing fiber) connections to the Purkinje cells (PC) of the cerebellar cortex (CC) and interpositus nucleus (IP). The Purkinje cells reciprocate inhibitory connections to the inferior olive through the IP. The IP nucleus provides descending motor control signals via the red nucleus (RN) and facial nucleus to elicit an eyeblink. In addition, the interpositus nucleus receives excitatory afferents from the pontine nuclei (PN), which themselves send ascending excitatory (mossy fiber) connections to granular (and Golgi) cells (GC) in the cerebellar cortex. These provide excitatory inputs to Purkinje cells (via parallel fibers). This connectivity architecture can be used to establish a function-structure mapping by assigning expectation and error units to the nuclei (and cerebellar cortex) to maximize the overlap between the anatomical connectivity and the connectivity implied by predictive coding. One such assignment is shown on the upper right, leading to the functional architecture on the left. (Left) This schematic depicts a functional anatomy implied by predictive coding. Here, we have associated exteroceptive, proprioceptive, and somatosensory prediction errors with the pontine, red, and IO nuclei, respectively. The hidden states (and their prediction errors) predicting spontaneous blinking, unconditioned stimuli, and conditioned stimuli have been assigned to the pontine nuclei, interpositus nucleus, and cerebellar cortex, respectively. Finally, the hidden states at the second level have been associated with the eyeblink center in the globus pallidus and the hippocampus (designated by PHS). The associated (second level) hidden causes have been placed in the pontine nuclei.
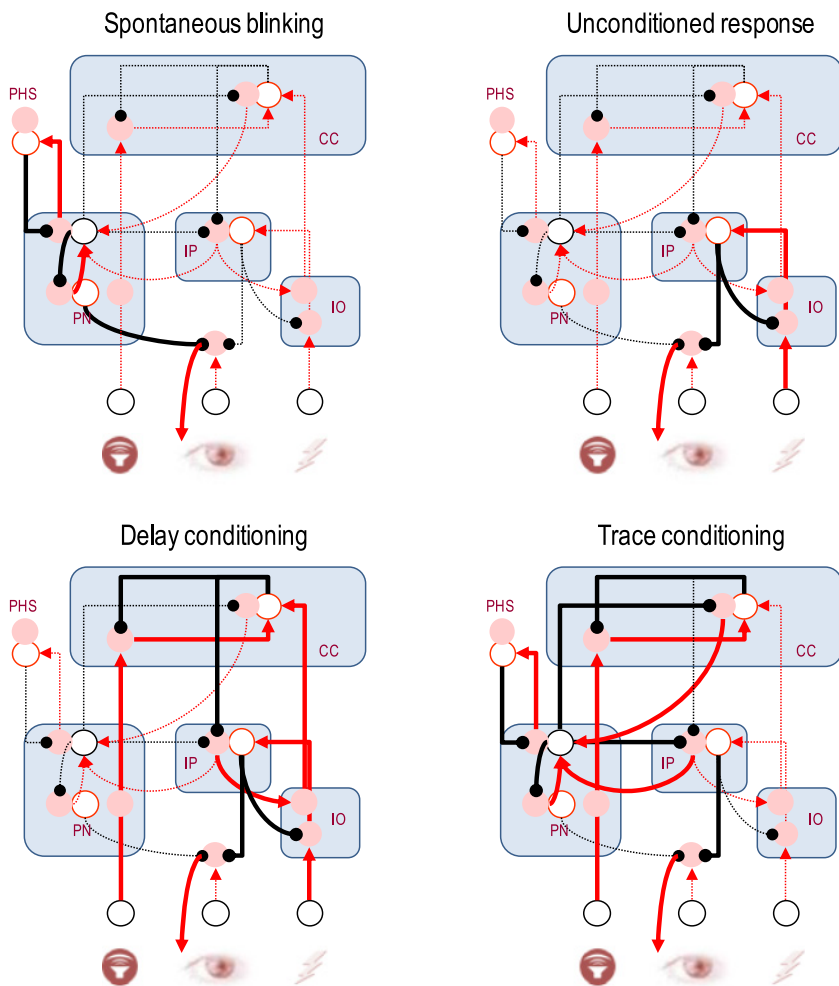
Purkinje cells in receipt of ascending prediction errors from the pontine nuclei (via mossy and parallel fibers) and those produced by expectations about an unconditioned stimulus from the interpositus and inferior olivary nuclei (via climbing fibers). After these expectations have been induced (by conditioned and unconditioned stimuli), they provide descending proprioceptive predictions to the output nuclei of the cerebellum (interpositus nucleus) and corollary discharge along parallel fibers. In short, delay conditioning engages reciprocal connections between the interpositus nucleus and cerebellar cortex.

Conversely, in trace conditioning, the unconditioned stimulus arrives after the termination of the conditioned stimulus. This calls on reciprocal

connections between the interpositus nucleus and the pontine nuclei. In this instance, expectations about the conditioned stimulus do not elicit proprioceptive predictions but reset the heteroclinic dynamics at the second level of the generative model, such that subsequent echo states or traces come to predict (and cause) an unconditioned response later in time. Notice that after conditioning, a conditioned response can occur in the absence of an unconditioned stimulus (which is why the ascending connections from the inferior olive have not been highlighted in the bottom-right panel of Figure 4). Having established a putative functional anatomy for blinking and eyeblink conditioning, we can now integrate or solve the predictive coding scheme to simulate unconditioned and conditioned responses.

Figure 4: This schematic highlights the different modes of blinking afforded by the predictive coding scheme of Figure 3. The four panels consider the key connections in spontaneous blinking, unconditioned responses to noxious stimuli, and conditioned responses under delay and trace conditioning. (Upper panels) Spontaneous blinking is mediated by descending predictions (from the eyeblink center in the globus pallidus) to the pontine nuclei and, finally, the red nucleus. Notice that there are no ascending prediction errors from the red nucleus. This is because the proprioceptive prediction error is resolved not through perceptual updating but by acting to fulfill the conditions (i.e., blinking). When blinking is elicited by an unconditioned stimulus, the descending proprioceptive predictions (or motor commands) descend from the interpositus nucleus. It can be seen (upper right) that these descending predictions target the red nucleus and inferior olive, providing motor commands and corollary discharge, respectively. (Lower panels) After learning or conditioning, further connections come into play to mediate a conditioned response. In delay conditioning, the unconditioned and conditioned stimuli are contemporaneous. It is therefore sufficient to expect contemporary exteroceptive and proprioceptive eyeblink signals. This calls on expectations of conditioned stimuli encoded by Purkinje cells that are in receipt of ascending prediction errors from the pontine nuclei (via mossy and parallel fibers) and those produced by expectations about an unconditioned stimulus from the interpositus an inferior olivary nuclei (via climbing fibers). When these expectations have been induced (by conditioned an unconditioned stimuli), they provide descending proprioceptive predictions to the output nuclei of the cerebellum (interpositus nucleus) and corollary discharge along parallel fibers. In short, delay conditioning engages reciprocal connections between the interpositus nucleus and cerebellar cortex. Conversely, in trace conditioning, the unconditioned stimulus arrives after the termination of the conditioned stimulus. This calls on reciprocal connections between the interpositus nucleus and the pontine nuclei. In this instance, expectations about the conditioned stimulus do not elicit proprioceptive predictions but reset the heteroclinic dynamics at the second level of the generative model, such that subsequent dynamical echo states or traces come to predict (and cause) an unconditioned response later in time.
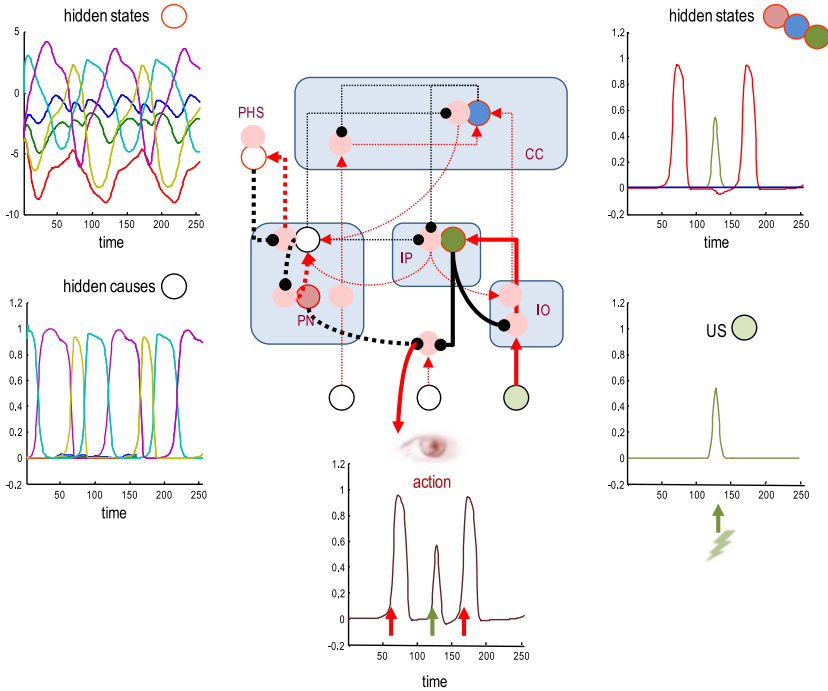
Figure 5: The results of a simulated unconditioned response. The central panel shows the predictive coding architecture, showing expectations about hidden states in different colors (red, spontaneous blinking; blue, conditioned stimulus; green, unconditioned response). The graphs show the activity of populations encoding expectations (and prediction errors) as a function of peristimulus time (expressed in time bins of about 250 ms). The lower right graph shows an unconditional stimulus, models with a gaussian bump function with a width of eight time bins. This produces somatosensory prediction errors (thick red connections) that induce an expected unconditioned response (green). This generates descending predictions (thick black connections), which elicit a blink (lower graph). The associated expectation is shown in green in the upper right graph. Here, we also see expectations about spontaneous blinking (red) generated by the heteroclinic cycle of hidden states (upper left graph) and their associated hidden causes (lower left graph). These mediate (spontaneous) eye blinks (red arrows) through the appropriate (dotted) connections.

5 Simulating Eyeblinks and Their Conditioning

Figure 5 shows the results of a simulated unconditioned response. These simulations used the active inference (predictive coding) scheme described

in the appendix and the generative model described in Figure 2. We assumed a low level of sensory noise with a log precision of 16. The log precision of the remaining random fluctuations was four unless otherwise stated. This intermediate precision was assigned to both levels of the hierarchy, thereby assigning equal precision to prediction error priors throughout the hierarchy. The middle panel shows the predictive coding architecture, highlighting expectations about hidden states in different colors (red, spontaneous blinking: blue, conditioned stimulus; green, unconditioned response). The graphs show associated activity in populations encoding expectations (and prediction errors) as a function of peristimulus time (in time bins of about 250 ms). The lower right graph shows an unconditional stimulus, modeled as a gaussian bump function with a width of eight time bins. This produces somatosensory prediction errors (thick red connections) that induce an expected unconditioned response (green). In turn, expectations about an unconditioned response generate descending predictions (thick black connections), which elicit a blink (bottom middle graph) that cancels the somatosensory prediction error. The associated expectation is shown in green in the upper right graph. Here, we also see expectations about spontaneous blinking (red) generated by the heteroclinic cycle of hidden states (top left graph) and their associated hidden causes (bottom left graph). These mediate (spontaneous) eyeblinks through the appropriate (dotted) connections, which occur at times indicated by the red arrows on the bottom middle graph. Note that there is no conditioned stimulus in this simulation, and therefore the unconditioned blink simply appears in addition to spontaneous blinking (and expectations about a conditioned stimulus are zero: blue line in the upper right graph). This should be contrasted with the responses shown in Figure 6.

Figure 6 contrasts predictive coding during spontaneous blinking, with the resetting of expectations by a conditioned stimulus such as a loud tone (before conditioning). The left panels show the dynamics of hidden states and causes as a function of peristimulus time during the emission of two spontaneous blinks. In this and subsequent figures, the first panel shows the sensory prediction (solid lines) and prediction error (dotted lines), although in this instance, the prediction errors are so small they are barely visible. The second (upper right) panel shows the expected hidden states generating sensory predictions using the same color scheme as in the previous figure. These hidden states are driven by a mixture of hidden causes shown on the lower left, which are softmax functions of the hidden states at the second level, shown on the lower right. During spontaneous blinking, the last hidden state and cause (olive green) generates expectations about blinking with a periodicity of about 120 time bins (or 3 seconds). The right panels show the same results but in response to a conditioned stimulus (e.g., tone) presented at around 125 time bins. In this case, there is a profound reorganization of the heteroclinic dynamics at the second level (lower right graph). This can be construed as a stimulus-dependent resetting of a central
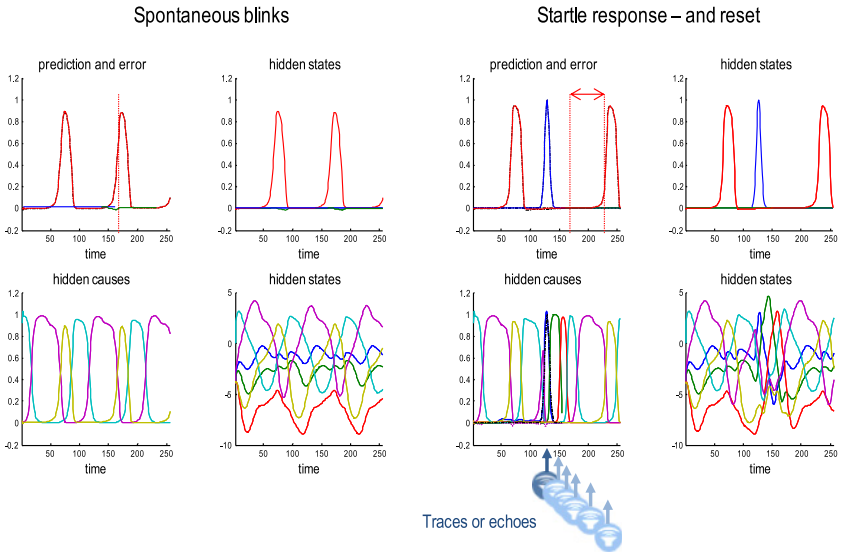
Figure 6: This figure contrasts predictive coding during spontaneous blinking with the resetting of second level (echo-state) expectations by a conditioned stimulus. (Left panels) These show the dynamics of hidden states and causes as a function of peristimulus time during the emission of two spontaneous blinks. In this and subsequent figures, the first panel shows the sensory prediction (solid lines) and prediction error (dotted lines). The second (upper right) panel shows the expected hidden states generating sensory predictions using the same color scheme as in the previous figure. These hidden states are driven by a mixture of hidden causes shown on the lower left, which are softmax functions of the hidden states at the second level, shown on the lower right. During spontaneous blinking, the last hidden state and cause (olive green) generates expectations about blinking every 120 time bins (or 3 seconds). (Right panels) These graphs show the same results but in response to a conditioned stimulus (e.g., tone) presented at around 125 time bins. In this case, there is a reorganization of the heteroclinic dynamics at the second level (lower right graph). This effectively delays the next spontaneous blink because the echo states have to pass through a sequence of unstable fixed points to return to the heteroclinic cycle. The delay (double-headed arrow) in spontaneous blinking reflects the delay before the heteroclinic cycle is resumed. This delay can be likened to a simple form of freezing or startle response.

pattern generator to produce echo states, starting with the first (blue line). This effectively delays the next spontaneous blink because the echo states have to pass through a sequence of unstable fixed points before entering the heteroclinic cycle again. This delay can be likened to a simple form of freezing or a startle response (Davis, Falls, Campeau, & Kim, 1993).

5.1 Simulations of Conditioning. The unconditioned responses, spontaneous blinking, and startle responses to conditioned stimuli above were all simulated before any conditioning. We repeated these simulations for several trials in which the conditioned stimulus (e.g., loud tone) and unconditioned stimulus (e.g., air puff) were presented together in delay (simultaneous presentation) or trace (sequential presentation) paradigms. The (free) parameters or connection strengths were initialized to zero and updated after every trial according to equation A.5 in the appendix.

Figure 7 shows the results of simulated delay conditioning. The graphs report expectations using the format of Figure 6, concentrating on a peristimulus time window of 64 time bins when the conditioned and unconditioned stimuli were presented. The conditioned stimulus (e.g., the tone) reached a maximum at 48 time bins (with a standard deviation of eight time bins). The unconditioned stimulus was presented toward the end of this period, reaching a maximum at 52 time bins (with a standard deviation of four time bins). The corresponding sensory inputs are shown as black dotted lines in the upper left graphs. The left panels show the responses prior to delay conditioning. Here, we see that the conditioned stimulus resets the echo states at the higher level (indicated by blue arrows), while the unconditioned stimulus elicits a proprioceptive prediction (and behavioral response) with an amplitude of about 0.5 arbitrary units. Note that the prediction of the unconditioned stimulus is only half of its actual amplitude, producing a rather large prediction error (dotted green line).

The right panels show the same responses after 16 trials of learning. At this point, the equivalent prediction error has been suppressed (and indeed shows a biphasic response with an anticipatory error). This reflects augmented expectations about an unconditioned stimulus furnished by the conditioned stimulus. Crucially, the ensuing behavioral response increases in amplitude with a decrease in latency (see the vertical lines in the upper left graph). We defined latency as a difference in onset of the eyeblink before and after conditioning (where onset time is the average obtained by treating the amplitude of the response as a probability distribution). The graph in the lower panel depicts the increase in response vigor (amplitude—broken line) and decrease in latency (solid line) over trials mediated by learning or plasticity (dotted lines). The learning or changes in parameters are shown in image format on the right to illustrate that conditioning is dominated by changes in the coupling between expectations about a conditioned stimulus (encoded by Purkinje cells in the cerebellar cortex) and consequent predictions about an unconditioned stimulus (in the interpositus nucleus). Interestingly, there is also an increase in the influence of (the first) hidden cause predicting the conditioned stimulus. This can be contrasted with the plasticity induced by trace conditioning shown in Figure 8.

Figure 8 shows the same simulation as Figure 7 but introducing a delay between the conditioned and unconditioned stimuli. Because learning is
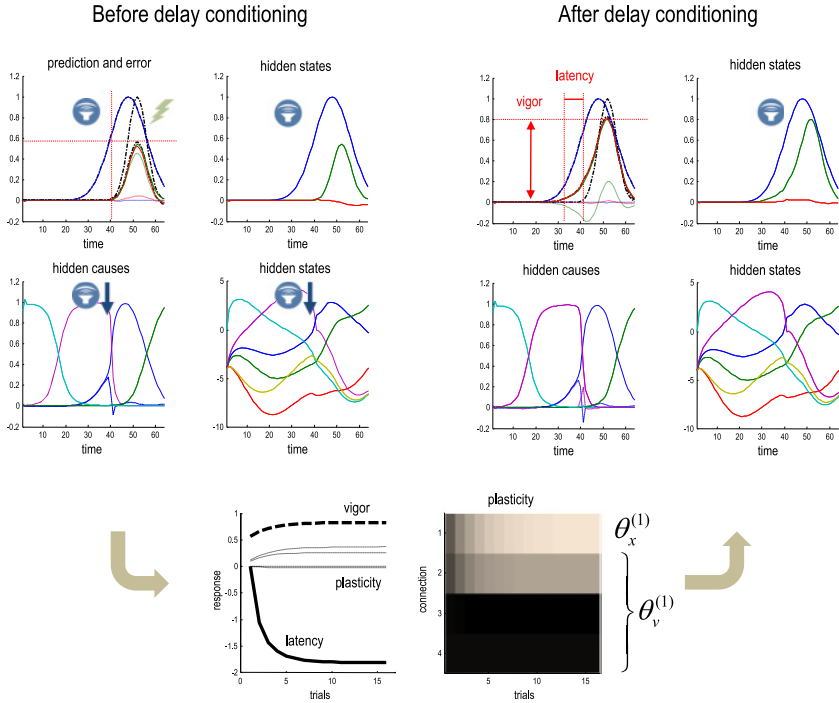
slightly slower in this trace conditioning paradigm, we doubled the number of trials to 32. The responses are shown over 128 time bins, where the peak of the unconditioned stimulus was delayed by 32 time bins. In this instance, response vigor increases more slowly, although latency is markedly reduced (at least for the first 20 or so trials, after which the latency reduction peaks). The key thing to note here is that ultimately, it is the connection from the third (echo or trace) hidden cause that mediates conditioning. This is because the delay means that this hidden cause is active at the time the unconditioned response is elicited. Crucially, this means a different set of connections is engaged—namely, the connections between the pontine nuclei and interpositus nucleus.

In summary, this section has established the face validity of the predictive coding scheme in relation to spontaneous blinking, startle responses to a stimulus prior to conditioning, and conditioned eyeblinks under both delay and trace conditioning. Furthermore, the ensuing computational anatomy matches the known circuitry and plasticity implicated in real eyeblink conditioning. In section 5.2, we try to establish the predictive validity of this scheme by trying to replicate the lesion deficit relationships observed empirically using simulated deafferentation.

5.2 Lesion-Deficit Simulations. There are many facts about the eyeblink conditioning system that have been disclosed through a long history of careful lesion studies (Davis et al., 1993; Freeman & Steinmetz, 2011; Kim et al., 1995; Yang et al., 2014). In brief, it is very well established that delay conditioning is most sensitive to lesions of the cerebellar output nuclei—namely, the interpositus nucleus. It is also clear that the cerebellar cortex plays a key role in delay (and trace) conditioning, although lesions to the cerebellar cortex per se are not always sufficient to block conditioning. Although not a double dissociation, there is an important dissociation between the conditioning deficits induced by cerebellar and hippocampal lesions. Hippocampal deafferentation destroys trace conditioning but leaves delay conditioning intact. Conversely, cerebellar lesions (especially of the interpositus nucleus) block both trace and delay conditioning. In what follows, we test the prediction that removing hippocampal afferents and cerebellar afferents would produce the same dissociation. In the simulated lesion studies, lesions were induced by reducing the log precision of selected error populations from 4 to 0. Effectively, this reduces their sensitivity or

---

Figure 7: An illustration of delay conditioning. The graphs report expectations using the format of Figure 6. Here, we concentrate on a peristimulus time window of 64 time bins, when the conditioned and unconditioned stimuli were presented. The conditioned stimulus reached a maximum at 48 time bins. The unconditioned stimulus was presented toward the end of this period, reaching a maximum at 52 time bins. The corresponding sensory inputs are shown as black dotted lines in the upper left graphs. (Left panels) These show the responses prior to delay conditioning. Here, we see that the conditioned stimulus resets the dynamics at the higher level, while the unconditioned stimulus elicits a proprioceptive prediction (and behavioral response) of about 0.5. Note that the prediction of the unconditioned stimulus is only half of its real amplitude, producing a rather large prediction error (dotted green line). (Right panels) After 16 trials of learning, this prediction error has been suppressed. This reflects augmented expectations about an unconditioned stimulus furnished by the conditioned stimulus. Crucially, the ensuing behavioral response increases in amplitude with a decrease in latency (see the vertical lines in the upper left graph). (Lower panel) The graph depicts the increase in response vigor (amplitude—broken line) and decrease in latency (solid line) over trials. The learning or changes in parameters mediating these behavioral changes are shown on the right to illustrate that conditioning is dominated by changes in the coupling between expectations about a conditioned stimulus (encoded by Purkinje cells in the cerebellar cortex) and consequent predictions about an unconditioned stimulus (in the interpositus nucleus). There is also an increase in the influence of (the first-) second-level hidden cause predicting the conditioned stimulus. In the simulations, we considered plasticity only in the contributions of the first three (second-level) hidden causes to an unconditioned response.
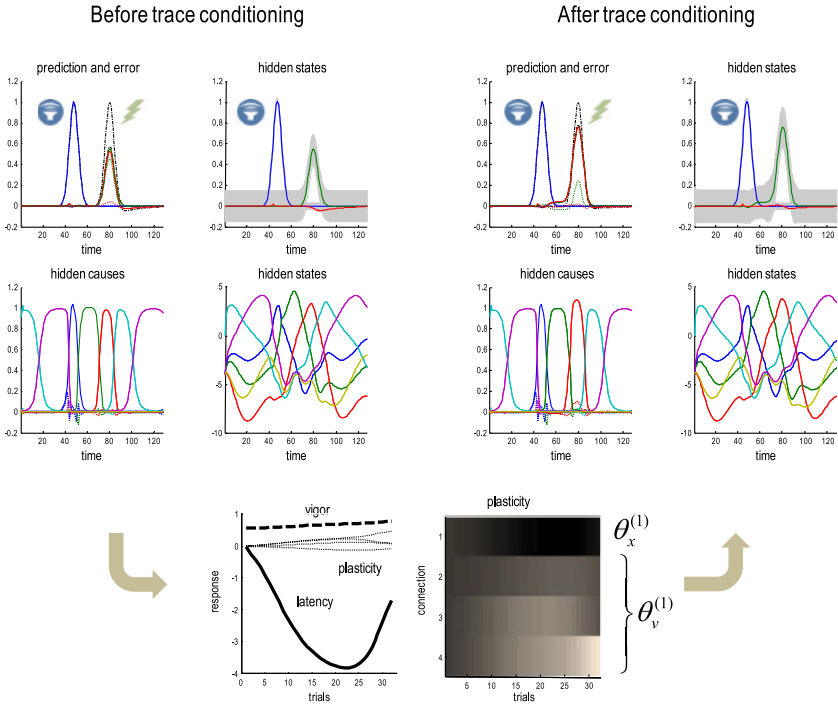
Figure 8: This figure uses the same format as Figure 7; however, we have now introduced a delay between the conditioned stimulus (e.g., tone) and unconditioned stimulus (e.g., shock) and doubled the number of trials to 32. These responses are shown over 128 time bins, where the peak of the unconditioned stimulus was delayed by 32 time bins. In this instance, response vigor increases more slowly, although latency is markedly reduced (at least for the first 20 or so trials). The key point to note here is that it is the connection from the third (echo or trace) hidden cause at the higher level that mediates conditioning.

gain and can be thought of as a reversible deactivation of specific neuronal populations.

Figure 9 shows the results of these simulated lesion experiments in terms of response latency and underlying plasticity. The upper rows show the results of conditioning under a variety of simulated experimental conditions—namely, normal learning, a deafferentation of hippocampal inputs, a lesion to the cerebellar output (interpositus) nuclei, and a lesion to the cerebellar cortex. The locations of these three forms of simulated deafferentation are shown in green in the bottom row. There are two key observations here. First is a marked dissociation between hippocampal and cerebellar lesions, where hippocampal lesions impair trace conditioning but
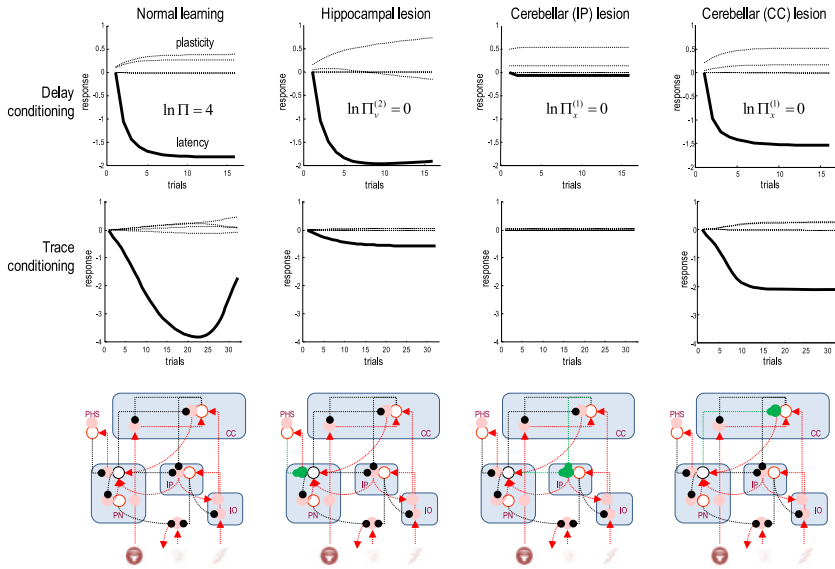
Figure 9: The results of simulated lesion experiments. The upper rows show the results of conditioning over 16 trials (for delay conditioning) or 32 trials (for trace conditioning) under a variety of simulated experimental conditions—namely, normal learning, a deafferentation of hippocampal inputs, a lesion to the cerebellar output (interpositus) nuclei, and a lesion to the cerebellar cortex. These lesions were implemented by reducing the gain or sensitivity of prediction error units in the respective regions from a log precision of 4 to 0. The locations of these three forms of simulated deafferentation are shown in green in the bottom row. There are two key observations here. First is a marked dissociation between hippocampal and cerebellar lesions, where hippocampal lesions impair trace conditioning but preserve delay conditioning. In contrast, a cerebellar lesion precludes both trace and delay conditioning. Second, lesions to the cerebellar output nuclei have more profound effects on conditioning to lesions of cerebellar cortex per se.

preserve delay conditioning. During delay conditioning, there is a failure to increase connectivity between the interpositus nucleus and the pontine nuclei (indeed this parameter goes slightly negative). However, these connections are not necessary for delay conditioning (see Figure 4). In contrast, a cerebellar lesion precludes both trace and delay conditioning. The second key observation is that lesions to the cerebellar output nuclei have more profound effects compared to lesions of cerebellar cortex per se. These observations are consistent with empirical findings.

Finally, we considered the interesting effects of lesions to the inferior olive on extinction. It is known that infusion of picrotoxin (which blocks
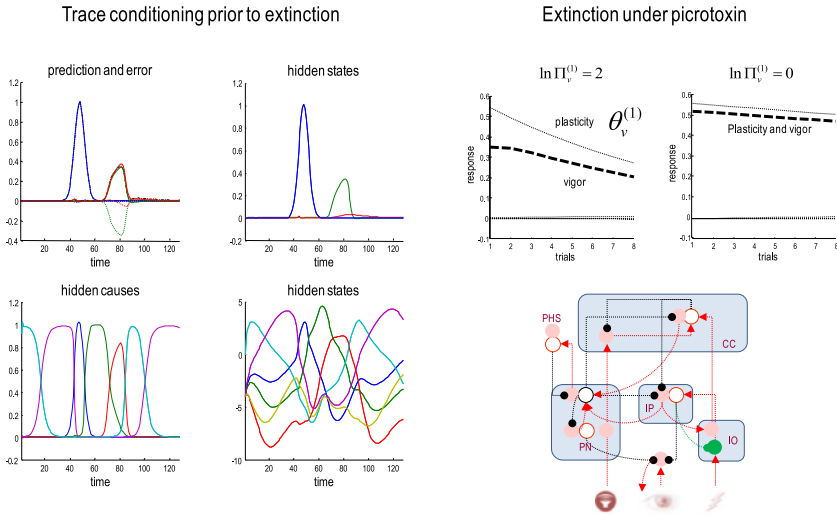
Figure 10: This figure simulates the effects of lesions to the inferior olive on extinction. (Left panel) These graphs report the results of a single trial after trace conditioning in the absence of an unconditioned stimulus. The format is the same as in Figure 8. Comparison with Figure 8 shows that the conditioned response (green lines in the upper left graph) is attenuated when not accompanied by an expected unconditioned stimulus. This simulation was repeated for eight trials to simulate extinction. (Right panel) The left graph shows extinction in terms of response vigor (dashed line), falling from the level of 0.35 (as in the left panels) to a value of about 0.2. This behavioral extinction is mediated by a decrease in the connectivity mediating trace conditioning (dotted lines). We then repeated the experiment but simulating a desensitization of the inferior olive responsible for reporting prediction errors about the unconditioned stimulus. Crucially, this simple lesion (reducing the log precision from 2 to 0) abolishes extinction.

the inhibitory neurotransmitter GABA) to the inferior olive prevents extinction (Medina, Nores, & Mauk, 2002). To examine the effect of a similar simulated lesion, we first had to reproduce extinction under the predictive coding scheme. These simulations are shown in Figure 10, where the left panels report the results of a single trial after trace conditioning but in the absence of an unconditioned stimulus. The format is the same as in Figure 8. Comparison with Figure 8 shows that the conditioned response (green lines in the upper left graph) is attenuated when not accompanied by an expected unconditioned stimulus. This simulation was repeated for eight trials to simulate extinction. The results are shown in the left graph

(right panel) in terms of response vigor (dashed line), falling from the level of 0.35 (as in the left panels) to a value of about 0.2.

This behavioral extinction is mediated by a decrease in the connectivity mediating trace conditioning (dotted lines). We then repeated the simulation but simulating a desensitization of principal cells in the inferior olive responsible for reporting prediction errors about the unconditioned stimulus. Crucially, this simple lesion (reducing the log precision from 2 to 0) abolishes extinction, as shown in the right graph (right panel). Note that this simulated chemical lesion also disinhibits the conditional response, which almost doubles in amplitude. In short, the simulated and empirical effects of (chemical) lesions to the inferior olive are consistent. The perspective afforded by predictive coding explains this effect simply: somatosensory prediction errors are attenuated, and therefore the system is unaware of the omission of an (expected) unconditioned stimulus. This means there are no subsequent prediction error signals to drive the plasticity necessary for extinction.

6 Conclusion

In summary, this letter provides an elemental account of learning and Pavlovian conditioning in terms of active inference and predictive coding. This scheme can account for spontaneous behavior, perceptual inference (and startle responses), and learning as disclosed through (delay and trace) conditioning. The implicit computational architecture maps very comfortably onto the known functional anatomy of eyeblink conditioning in cerebellar circuits and associated (brain stem) nuclei. Furthermore, we were able to establish the predictive validity of the ensuing structure-function mapping using simulated lesion experiments that reproduce empirical findings remarkably well. Specifically, we were able to reproduce the dissociation between deficits in delay and trace conditioning produced by lesions to the hippocampus and cerebellar systems, respectively.

Clearly, there are many details of neuronal implementation and a vast number of empirical findings that we have not covered; however, the minimal description on offer has some interesting implications for understanding perception in the context of learning and behavior. One interesting aspect is that active inference can explain various forms of (Pavlovian) learning without reference to reinforcement learning. This reflects the fact that standard notions from reinforcement learning (e.g., value functions) can always be cast in terms of prior beliefs, which are invariably part of a generative model. This has the advantage that the same imperatives (and process theories) can be applied identically to perception and action.

Furthermore, the hierarchical nature of (active or embodied) inference speaks to a potentially important relationship between various forms of

learning and control. For example, the simulations above could be regarded as successive hierarchical elaborations of beliefs about behavior. In other words, delay conditioning can be understood as the contextualization of an unconditioned response (to an unconditioned stimulus) through the learning of contingencies relating to conditioned stimuli. This learning invokes cerebellar representations of conditioned stimuli that jointly predict the conditioned stimulus and unconditioned response. In a similar vein, trace conditioning can be regarded as a hierarchical extension of this internal modeling, in which temporally extensive dynamics at higher hierarchical levels contextualize lower-level expectations that are distributed over time. This may reflect a fundamental aspect of learning—namely, the emergence of habitual or Pavlovian behavior under deep hierarchical inference about the causes of our sensations (Pezzulo, Rigoli, & Chersi, 2013).

Appendix: Generalized Predictive Coding and Active Inference ‗‗‗‗‗‗‗

This brief description of generalized predictive coding is based on Friston (2008). This scheme is based on three assumptions:

- The brain minimizes a free energy of sensory inputs defined by a generative model.
- The generative model used by the brain is hierarchical, nonlinear, and dynamic.
- Under this model, neuronal firing rates encode the expected state of the world.

Free energy is a quantity from statistics that measures the quality of a model in terms of the probability that it could have generated observed outcomes. This means that minimizing free energy maximizes the Bayesian evidence for the generative model. The second assumption is motivated by noting that the world is both dynamic and nonlinear and that hierarchical causal structure emerges inevitably from a separation of spatial and temporal scales. The final assumption is the Laplace assumption, which leads to a simple and flexible neuronal code.

Given these assumptions, one can simulate a whole variety of neuronal processes by specifying the particular equations that constitute the brain's generative model. In brief, these simulations use differential equations that minimize the free energy of sensory input using a generalized gradient descent:

$$\dot{\tilde{\mu}}(t) = D\tilde{\mu}(t) - \partial_{\tilde{\mu}} F(\tilde{s}(t), \tilde{\mu}(t)). \tag{A.1}$$

These differential equations say that neuronal activity encoding posterior expectations about (generalized) hidden states of the world $\tilde{\mu} = (\mu, \mu', \mu'', \dots)$ reduces free energy—where free energy $F(\tilde{s}, \tilde{\mu})$ is a function of sensory inputs $\tilde{s} = (s, s', s'', \dots)$ and neuronal activity. This is known

as generalized predictive coding or Bayesian filtering. The first (prediction) term of the predicted changes in expectations is based on a differential matrix operator D that returns their generalized motion $D\tilde{\mu} = (\mu', \mu'', \mu''', \ldots)$. The second (correction) term is usually expressed as a mixture of prediction errors that ensures that the changes in posterior expectations are Bayes-optimal predictions about hidden states of the world. To perform neuronal simulations under this scheme, it is only necessary to integrate or solve equation A.1. This solution simulates the neuronal dynamics that encode posterior expectations. These expectations depend on the brain's generative model of the world, which we assume has the following hierarchical form:

$$s = g^{(1)}(x^{(1)}, v^{(1)}) + \omega_v^{(1)}$$
$$\dot{x}^{(1)} = f^{(1)}(x^{(1)}, v^{(1)}) + \omega_x^{(1)}$$
$$\vdots$$
$$v^{(i-1)} = g^{(i)}(x^{(i)}, v^{(i)}) + \omega_v^{(i)}$$
$$\dot{x}^{(i)} = f^{(i)}(x^{(i)}, v^{(i)}) + \omega_x^{(i)}$$
$$\vdots \tag{A.2}$$

This equation describes a probability density over the sensory and hidden states that generate sensory input. Here, the hidden states have been divided into hidden states and causes $(x^{(i)}, v^{(i)})$ at the ith level within the hierarchical model. Hidden states and causes are abstract variables that the brain uses to explain or predict sensations, like the motion of an object in the field of view. Hidden causes link hierarchical levels, whereas hidden states link dynamics over time.

The generative model is characterized by nonlinear functions $(f^{(i)}, g^{(i)})$ of hidden states and causes that generate hidden causes for the level below and, at the lowest level, sensory inputs. Random fluctuations in the motion of hidden states and causes $(\omega_x^{(i)}, \omega_v^{(i)})$ enter each level of the hierarchy. Gaussian assumptions about these random fluctuations make the model probabilistic. In other words, they play the role of sensory noise at the first level and induce uncertainty at higher levels. The amplitudes of these random fluctuations are quantified by their precisions $(\Pi_x^{(i)}, \Pi_v^{(i)})$.

A.1 Perception. Given the form of the generative model, equation A.2, we can now write the differential equations, A.1, describing neuronal dynamics in terms of (precision-weighted) prediction errors. These errors represent the difference between posterior expectations and predicted values under the generative model (using $A \cdot B \triangleq A^T B$ and omitting higher-order

terms):

$$\dot{\tilde{\mu}}_x^{(i)} = D\tilde{\mu}_x^{(i)} + \frac{\partial \tilde{g}^{(i)}}{\partial \tilde{\mu}_x^{(i)}} \cdot \xi_v^{(i)} + \frac{\partial \tilde{f}^{(i)}}{\partial \tilde{\mu}_x^{(i)}} \cdot \xi_x^{(i)} - D^T \xi_x^{(i)},$$

$$\dot{\tilde{\mu}}_v^{(i)} = D\tilde{\mu}_v^{(i)} + \frac{\partial \tilde{g}^{(i)}}{\partial \tilde{\mu}_v^{(i)}} \cdot \xi_v^{(i)} + \frac{\partial \tilde{f}^{(i)}}{\partial \tilde{\mu}_x^{(i)}} \cdot \xi_x^{(i)} - \xi_v^{(i+1)}.$$

$$\xi_x^{(i)} = \Pi_x^{(i)} \cdot \tilde{\varepsilon}_x^{(i)},$$

$$\xi_v^{(i)} = \Pi_v^{(i)} \cdot \tilde{\varepsilon}_v^{(i)}.$$

$$\tilde{\varepsilon}_x^{(i)} = D\tilde{\mu}_x^{(i)} - \tilde{f}^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)}),$$

$$\tilde{\varepsilon}_v^{(i)} = \tilde{\mu}_v^{(i-1)} - \tilde{g}^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)}). \tag{A.3}$$

This produces a relatively simple update scheme, in which posterior expectations $\tilde{\mu}^{(i)}$ are driven by a mixture of prediction errors $\tilde{\varepsilon}^{(i)}$ that are defined by the equations of the generative model.

In neuronal network terms, equation A.3 says that error units compute the difference between expectations at one level and predictions from the level above to produce a (precision-weighted) prediction error $\xi^{(i)}$ at the ith level of the hierarchy. Posterior expectations are then driven by prediction errors from the same level and the level below. These constitute bottom-up and lateral messages that drive expectations to produce better predictions and reduce prediction error in the level below. In the cortex, the sources of bottom-up prediction errors are generally thought to be superficial pyramidal cells, because they send forward (ascending) connections to higher cortical areas. Conversely, predictions are thought to be conveyed from deep pyramidal cells by backward (descending) connections, to target the superficial pyramidal cells encoding prediction error (Bastos et al., 2012; Mumford, 1992).

A.2 Learning. The corresponding updates for the parameters follow a similar scheme; however, because parameters change slowly, we have a simpler gradient descent that effectively accumulates free energy gradients over time:

$$\dot{\tilde{\mu}}_\theta(t) = -\partial_{\tilde{\mu}} S(\tilde{s}, \tilde{\mu}),$$

$$\partial_t S = F(\tilde{s}(t), \tilde{\mu}(t)). \tag{A.4}$$

Here, $S(\tilde{s}, \tilde{\mu})$ is the antiderivative of free energy (known as free action). Neurobiologically, when expressed in terms of prediction errors, the solution

to equation A.4 corresponds to associative plasticity. For parameters that affect the equations of motion, we have (with a formally identical equation for parameters that affect the nonlinear mappings)

$$\dot{\mu}_{ij}^{\theta} = \alpha_{ij}^{\theta} - \Pi_{ij}^{\theta} \varepsilon_{ij}^{\theta},$$

$$\dot{\alpha}_{ij}^{\theta} = -\frac{\partial \tilde{\varepsilon}_x^{(i)}}{\partial \mu_{ij}^{\theta}} \cdot \Pi_x^{(i)} \cdot \tilde{\varepsilon}_x^{(i)}. \tag{A.5}$$

This says that plasticity comprises an associative term and a decay term mediating priors on the parameters (first equality). The integral of the associative term is simply the covariance between presynaptic input and postsynaptic prediction error (second equality); in short, it mediates associative or Hebbian plasticity. Crucially, note that the plasticity mediated by a positive prediction error is negative. In other words, from the perspective of presynaptic inputs to principal cells encoding (postsynaptic) prediction error, this would manifest as synaptic depression of the sort seen in the cerebellum.

A.3 Action. In active inference, expectations elicit behavior by sending top-down predictions down the hierarchy that are unpacked into proprioceptive predictions at the level of the cranial nerve nuclei (e.g., facial nucleus) and spinal cord. These engage classical reflex arcs to suppress proprioceptive prediction errors and produce the predicted motor trajectory:

$$\dot{a} = -\frac{\partial F}{\partial a} = -\frac{\partial \tilde{s}}{\partial a} \cdot \xi_v^{(1)}. \tag{A.6}$$

The reduction of action to classical reflexes follows because the only way that action can minimize free energy is to change sensory (proprioceptive) signals (see the equilibrium point formulation of motor control: Feldman & Levin, 1995). In short, active inference can be regarded as equipping a generalized predictive coding scheme with classical reflex arcs.

Acknowledgments ――――――――――――――――――――――――――

References ――――――――――――――――――――――――――――

Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Struct. Funct.*, *218*(3), 611–643.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.*, *36*(3), 181–204.

Davis, M., Falls, W. A., Campeau, S., & Kim, M. (1993). Fear-potentiated startle: A neural and pharmacological analysis. *Behav. Brain Res.*, *58*(1–2), 175–198. doi:10.1016/0166-4328(93)90102-v

Dayan, P., Hinton, G. E., & Neal, R. (1995). The Helmholtz machine. *Neural Computation*, *7*, 889–904.

Feldman, A. G., & Levin, M. F. (1995). The origin and use of positional frames of reference in motor control. *Behav. Brain Sci.*, *18*, 723–806.

Freeman, J. H., & Steinmetz, A. B. (2011). Neural circuitry and plasticity mechanisms underlying delay eyeblink conditioning. *Learn. Mem.*, *18*(10), 666–677. doi:10.1101/lm.2023011

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput Biol.*, *4*(11), e1000211.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris*, *100*(1–3), 70–87.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biol. Cybern.*, *104*, 137–160.

Friston, K. J., Shiner, T., Fitzgerald, T., Galea, J. M., Adams, R., Brown, H., . . . Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS Comput. Biol.*, *8*(1), e1002327.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proc. R. Soc. Lond. B*, *171*, 179–196.

Helmholtz, H. (1866/1962). *Treatise on physiological optics* (vol. 3). (J. Southall, Trans.) New York: Dover.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Kim, J. J., Clark, R. E., & Thompson, R. F. (1995). Hippocampectomy impairs the memory of recently, but not remotely, acquired trace eyeblink conditioned responses. *Behavioral Neuroscience*, *109*(2), 195–203. doi:10.1037/0735-7044.109.2.195

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. Opt. Image Sc. Vis.*, *20*, 1434–1448.

McCormick, D. A., & Thompson, R. F. (1984). Cerebellum: Essential involvement in the classically-conditioned eyelid response. *Science*, *223*(4633), 296–299. doi:10.1126/science.6701513

Medina, J. F., Nores, W. L., & Mauk, M. D. (2002). Inhibition of climbing fibres is a signal for the extinction of conditioned eyelid responses. *Nature*, *416*(6878), 330–333. doi:10.1038/416330a

Mumford, D. (1992). On the computational architecture of the neocortex. II. *Biol. Cybern.*, *66*, 241–251.

Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Front. Psychol.*, *4*, 92. doi:10.3389/fpsyg.2013.00092

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, *2*(1), 79–87.

Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B. Biol. Sci.*, *216*(1205), 427–459.

Tani, J., & Nolfi, S. (1999). Learning to perceive the world as articulated: An approach for hierarchical learning in sensory–motor systems. *Neural Netw.*, *12*, 1131–1141.

Verschure, P. F., Voegtlin, T., & Douglas, R. J. (2003). Environmentally mediated synergy between perception and behavior in mobile robots. *Nature*, *425*, 620–624.

Yang, Y., Lei, C., Feng, H., & Sui, J. F. (2014). The neural circuitry and molecular mechanisms underlying delay and trace eyeblink conditioning in mice. *Behav. Brain Res.*, *278c*, 307–314. doi:10.1016/j.bbr.2014.10.006