Thanks for the comments. There were some suggested improvements – if accepted we will incorporate: an algorithm box; analysis of complexity; redesign and/or annotation for Figure 1; Figure 2A adaptation; and typo corrections.

R1: Fig 2: We will delete the reference to a gray line, which had been removed from the figure – sorry. Parameters t_0 and τ affect readout, not the internal model or inference: t_0 is the perceived time of the flash; τ is the postdictive window – the duration of subsequent evidence that contributes to the estimate of the location of the moving object at t_0.
Comparisons: Biological plausibility is at the heart of the problem we seek to address. Postdictive behaviour depends on updating beliefs about *past* states – the central question of our work is how this might happen in natural online inference. Standard temporal variational methods (e.g. Chung et al., 2015; Fraccaro et al., 2017) represent states individually, and so would require acausal ("backward") message passing to revise past states (see also comments under Signatures below). This is possible (e.g. Fraccaro; or Johnson et al., 2016), but currently only implemented for restricted classes of generative model which do not capture the natural environment.

R2: Encoding functions: indeed, we chose fixed but random $\boldsymbol{\gamma}$. Results are robust to redrawing random projections, and to different nonlinearities. To train the parameters of $\boldsymbol{\gamma}$ and $\boldsymbol{\psi}_t$ one could minimize the predictive error (5) or (9) wrt these parameters. Gradient descent would be a natural choice, albeit with debatable biological plausibility.
Signatures of internal representation: a general attempt to distinguish encoding schemes that use expectations (DDC), samples, or natural parameters (including the 'PPC') depends on additional model assumptions and is larger than can be addressed satisfactorily in a NeurIPS paper. However, in the specific context of temporal models and postdiction, these schemes differ in the mechanism by which beliefs about the past may be maintained and updated. The linearity of expectation, combined with temporal encoding functions (or sufficient statistics) $\boldsymbol{\psi}_t$, allows a simple dynamical update rule to maintain and update beliefs (eqs (7) and (10) or (11)). If beliefs about latent states were represented by samples, then these samples would need to be maintained in a sort of "buffer" to be modified by later inputs and accessed by downstream processing. Natural parameter encodings could, in principle, exploit temporal sufficient statistic functions similar to those we introduce, but the dynamical updates required to maintain and revise the natural parameters through time would be far more complex to derive and implement than for the mean parameters of the DDC. Thus, of these options, we believe the DDC inference and postdiction in the temporal context comes closest to the known biology.
Lines 153-154: (We will clarify the entire first paragraph of 3.2 if accepted.) Equation (8) would have been better written: $\mathcal{L}^f(\widetilde{\boldsymbol{W}}_t; \boldsymbol{x}_{1:t\text{-}1}) = \mathbb{E}_{q(\boldsymbol{z}_{1:t}, \boldsymbol{x}_t | \boldsymbol{x}_{1:t\text{-}1})}\left[\|\widetilde{\boldsymbol{W}}_t \boldsymbol{\sigma}(\boldsymbol{x}_t) - \boldsymbol{\psi}(\boldsymbol{z}_{1:t})\|_2^2\right]$. That is, the loss is history-dependent and the optimal weights thus depend on both history and time. Now, if we restrict functions $h_{\boldsymbol{W}}$ in (9) to be linear in $\boldsymbol{\sigma}(\boldsymbol{x}_t)$ (as both (10) and (11) are) then $h_{\boldsymbol{W}}$ can be seen as effectively mapping $\boldsymbol{r}_{t\text{-}1}$ to a time- and history-dependent $\widetilde{\boldsymbol{W}}_t$. Thus, although the loss (9) is the expectation of (8) over histories, the optimal sequence of $\widetilde{\boldsymbol{W}}_t$ for each history still provides a target for the minimisation of (9), and the expectation of the minima of (8) bounds the minimum of (9) below. But indeed, the constrained form of $h_{\boldsymbol{W}}(\boldsymbol{r}_{t\text{-}1}, \cdot)$ means that this lower bound will not generally be achieved.

R3: "Neurons": Our intention is for activities of neurons in our scheme to provide a model for the firing rates of biological neurons. We expect DDC-based computations to be robust to Poisson-like noise, as linear operations acting on populations of neurons (inference and readout) effectively average away independent perturbations. Consistent with this intuition, when we re-ran the flash-lag experiment using Poisson neurons the mean results were essentially the same, albeit with greater trial-to-trial variance. For occluded tracking, the R^2 in postdicting the true $\boldsymbol{z}_t$ by posterior mean still increased with postdictive window τ, although this improvement did not extend as far into the past as in the noiseless case. If accepted, we will include these new experimental results.
Line 92: Indeed, DDC activities encode a *distribution* or *belief* about a random variable – we will amend the text.
Line 118: Our focus here is to propose how neural circuits may implement postdictive inference – itself, a non-trivial problem even when the internal model of the world is known – and relate such inference to psychometric phenomena. A general approach to modelling learning using DDC-based inference in wake-sleep has been proposed before (Vertes and Sahani, 2018), although extensions to temporal models are the subject of current research.
Learning the readout: the key feature of (3) and (12) is that $\boldsymbol{\alpha}$ can be found for known target functions using evaluations of [e.g. for (12)] $\boldsymbol{\psi}(\boldsymbol{z}_{1:t})$ and $l(\boldsymbol{z}_{t\text{-}\tau})$, and distributional uncertainty then propagates appropriately. Thus (as with the recognition weights) we need only simulations of $\boldsymbol{z}_{1:t}$ from the internal model, potentially from the sleep phase of wake-sleep. More generally, the target function $l()$ might need to be learnt based on supervision or reinforcement. This could follow "normal" learning rules; with the DDC again ensuring that uncertainty is handled correctly.
Lines 190-195: These simulations probe the quality of inference in the learnt model. Figure 1D reproduces the known effect that perceptual continuity depends on the level of the interrupting noise. Figure 1E demonstrates quantitative inference – the posterior on continuity is reduced when the noise is quieter still. Finally, 1F demonstrates that the inference model has correctly encoded the fact that tones in the generative model do not change in loudness.
Fig 1, multiple buses: The hallmark of postdiction is that beliefs about past values are revised by new evidence. Thus, we show beliefs at time step i (column) conditioned on sensory experience up to various later times j (row). That is, the three boxes in the ith column of the jth row show decoded posterior probabilities for the three possible levels of the tone at step j, based on sensory input up to time i. The smaller inset buses show the inferred noise level in the same way.