Stubborn predictions in primary visual cortex

Daniel Yon[1,2], Emily R Thomas[1,3], Sam J Gilbert[4], Floris P de Lange[5], Peter Kok[6] & Clare Press[1,6]

1. Department of Psychological Sciences, Birkbeck, University of London, UK.
2. Department of Psychology, Goldsmiths, University of London, UK.
3. Neuroscience Institute, New York University Medical Center, USA.
4. Institute of Cognitive Neuroscience, University College London, UK.
5. Donders Institute for Brain, Cognition and Behaviour, Radboud University, NL.
6. Wellcome Centre for Human Neuroimaging, Queen Square Institute of Neurology, University College London, UK.

Correspondence: d.yon@bbk.ac.uk

Accepted at *Journal of Cognitive Neuroscience* on 30th March 2023.

Conflict of interest statement: The authors declare they have no competing interests.

Abstract

Perceivers can use past experiences to make sense of ambiguous sensory signals. However, this may be inappropriate when the world changes and past experiences no longer predict what the future holds. Optimal learning models propose that observers decide whether to stick with or update their predictions by tracking the uncertainty or 'precision' of their expectations. But contrasting theories of prediction have argued that we are prone to misestimate uncertainty – leading to stubborn predictions that are difficult to dislodge. To compare these possibilities, we had participants learn novel perceptual predictions before using fMRI to record visual brain activity when predictive contingencies were disrupted - meaning that previously 'expected' events become objectively improbable. Multivariate pattern analyses revealed that expected events continued to be decoded with greater fidelity from primary visual cortex, despite marked changes in the statistical structure of the environment which rendered these expectations no longer valid. These results suggest that our perceptual systems do indeed form stubborn predictions even from short periods of learning – and more generally suggest that top-down expectations have the potential to help or hinder perceptual inference in bounded minds like ours.

Introduction

Perceiving creatures need to generate faithful representations of the external world from noisy and ambiguous signals. Observers can overcome this inherent sensory ambiguity by combining incoming sensory data with prior knowledge about what the world is likely to contain (de Lange et al., 2018; Heilbron & Chait, 2018; Hogendoorn, 2022; Yuille & Kersten, 2006). For example, 'sharpening' models hypothesise that we use top-down predictions to relatively upweight the activity of neurons tuned to sensory features we expect, via competitive interactions that suppress populations tuned to unexpected alternatives (see Figure 1). We have observed this kind of predictive retuning in early and late visual brain areas (Kok et al., 2012; Yon et al., 2018), with suppression of unexpected signals reshaping neural activity so that it more closely resembles our prior predictions (see also González-García & He, 2021).

Predictively retuning sensory populations could provide an adaptive way to deal with sensory ambiguity in stable environments, as different kinds of retuning can weight our perceptual systems towards events that we expect - which are more likely to occur. However, a system that uses predictions to finesse ambiguous inferences encounters a problem when the world

begins to change, as predictions based on the past may cease to be useful when interpreting new environments.

In a labile world like ours, perceivers must thus decide whether to stick with their existing expectations or learn and update their models in the face of new data. Computational models of learning suggest an optimal solution to this dilemma can be found if agents estimate the 'precision' or reliability of their predictions and the stability of the outside world – updating beliefs more when our most reliable predictions are violated and when we think our environments are more changeable (Behrens et al., 2007; O'Reilly, 2013; Yu & Dayan, 2005). Faithfully estimating these 'parameters' of our internal models and our external world can optimise how we use and update our predictions, ensuring that we do not bring old expectations to bear on new situations where they no longer apply.

However, many modern models assume that our estimates of 'precision' or uncertainty in our predictions often decouple from reality (Yon & Frith, 2021). For example, contemporary predictive coding models incorporate the notion of 'stubborn predictions' – assuming that some top-down expectations are difficult to update in the face of new information (Yon et al., 2019). In model-based terms, stubborn predictions are assumed to arise when agents assign especially high 'precision' to top-down predictions at the expense of ascending error signals that could potentially update them (Friston, 2018).

Though prior work has investigated how learned expectations influence activity in frontal and parietal decision circuits (e.g., Hansen et al., 2011) at present we do not know whether our perceptual systems form stubborn predictions that shape activity in the sensory brain, or whether our learned expectations are updated appropriately when the environment signals they are no longer reliable. Here, we compare these possibilities. Our participants learned novel predictive associations between executed actions (finger movements) and the identity of ensuing visual stimuli (oriented gratings). After training with perfectly deterministic contingencies, we recorded participants' visual brain activity with 3T fMRI during the same experimental task but with previously learned relationships disrupted: expected outcomes, unexpected outcomes or omitted outcomes were experienced with equal probability (33.3%). Optimal theories of learning would suggest that the shift to a new context where predictions are more often violated than confirmed should drive model updating – abandoning predictions that are no longer reliable. However, if perceptual systems acquire predictions that are suboptimally stubborn, expectations established in previous contexts may not be dislodged by discrepant experiences – and observers may still rely on old predictions in new settings, with detectable influences on sensory brain activity.

In line with this latter possibility, we found that patterns of brain activity in primary visual cortex showed the same influences of prediction identified in previous work (Kok et al., 2012; Yon et al., 2018), even though 'expected' events were objectively improbable (i.e., observers were more likely to encounter an 'unexpected' or omitted outcome, rather than the outcome they had previously learned to expect). This suggests that even short periods of learning can lead us to acquire somewhat stubborn predictions, which bias representations in the sensory brain – even though such biases will not improve perceptual inference. These results suggest that human perceivers may misestimate the reliability of their predictions, relying on past learning even when the world signals that relationships have changed.

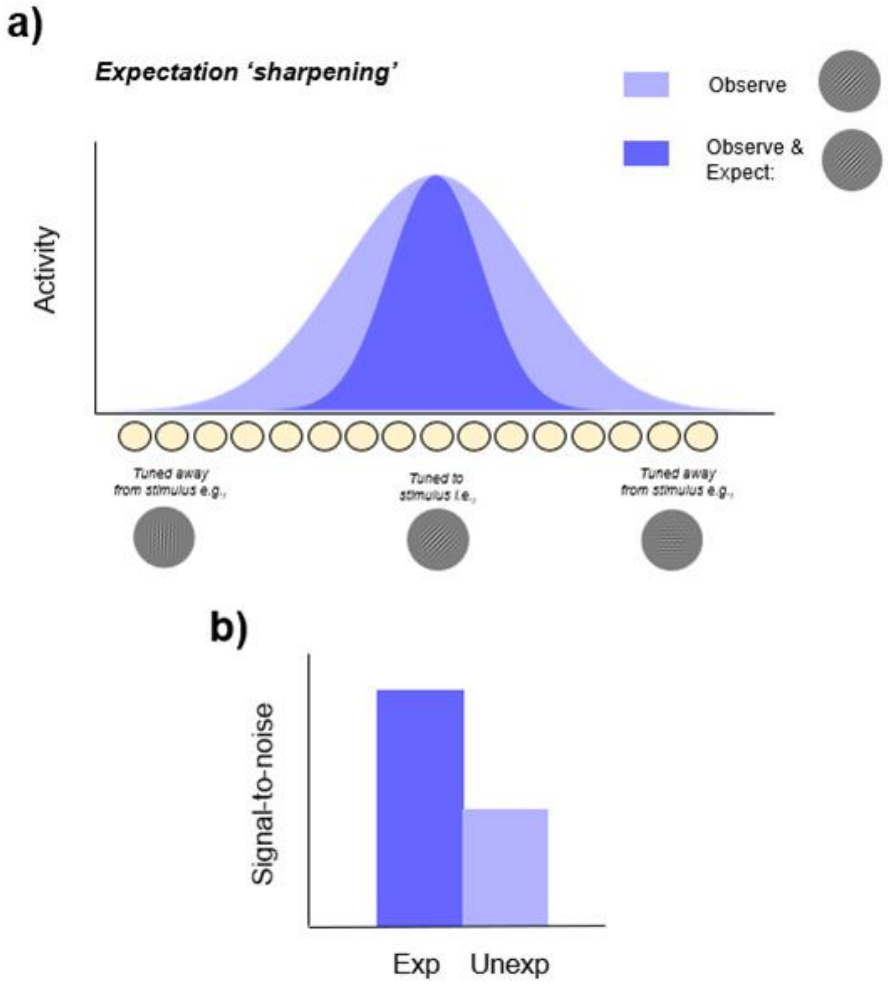

Figure 1: Expectation 'sharpening' is one mechanism proposed in the literature that could generate superior decoding of expected sensory events. Sharpening models of perceptual prediction assume that probabilistic knowledge reshapes activity in sensory brain areas, such that it is weighted more heavily towards prior expectations (de Lange et al., 2018). **a)** For example, in a population of visual neurons, an observed stimulus (e.g., clockwise-tilted grating) will evoke most activity in those units tuned to the observed (clockwise tilt) features and lesser activity in those tuned to other features (e.g., counter-clockwise tilt). According to sharpening accounts, top-down predictions act to relatively upweight expected sensory signals by suppressing activity in populations *tuned away* from the stimulus (Press & Yon, 2019). **b)** This kind of predictive 'sharpening' (i.e. suppression of unexpected noise) leads to a greater signal-to-noise ratio across the population when expectations are valid. Thus, the presence of sharpening can be evaluated empirically using multivariate neuroimaging techniques (e.g., decoding) to quantify how expectations change the information content of sensory brain regions.

Methods

Participants: Twenty one right-handed participants were recruited (13 female, mean [SD] age = 27.9 [5.5] years) from Birkbeck, University of London and University College London (UCL). All participants reported normal or corrected to normal vision and had no history of psychiatric or neurological illness. One participant was excluded from the sample due to

excessive movement during scanning (i.e. sudden movements >2mm throughout runs), leaving a final simple of 20. The experiment was approved by local ethics committees at Birkbeck and UCL. Sample size was determined based on previous studies which have observed reliable effects of expectation on multivariate measures of visual brain activity (Kok et al, 2012; Yon et al, 2018).

Apparatus: Experimental task and stimuli were generated using Cogent in Matlab. During training, stimuli were displayed on a grey background via a Dell Laptop 14" LCD screen (60 Hz). During scanning the same stimuli were displayed on a rear-projection screen using a JVC DLA-SX21 projector (60 Hz). In both cases, stimuli subtended ~15° visual angle. Participants registered their perceptual decisions (see below) using an MRI compatible button box, and used the same button box when performing the task in and outside of the scanner.

Experimental procedure: Participants completed a combined action and perceptual decision-making task. Each trial began with the presentation of a central fixation cross framed by an imperative cue (square or triangle), which instructed participants to abduct either their right index or little finger. Participants' abduction movements (i.e., releasing a previously depressed button) triggered the presentation of an oriented grating stimulus for 500 ms. Each stimulus was a sinusoidal grating, enveloped by a Gaussian filter to create Gabor patches of 80% Michelson contrast, at 1.5 cycles/ °. Stimuli were either oriented clockwise (CW; 45°) or counter-clockwise (CCW; 135°) relative to the vertical midline, and appeared within an annulus such that the central fixation cross remained visible. After a variable 300 – 500 ms delay, participants were asked to judge the orientation of the observed stimulus. Participants were presented with a question probing a specific orientation (e.g., "Was the stimulus tilted CW?"), and both orientations were probed with equal probability. They responded 'yes' or 'no' using their left thumb. Participants were required to respond within 1500 ms, and the next trial started after a variable inter-trial interval (2-3 s in training and 2-6 s in the scanner, see Figure 2 below).

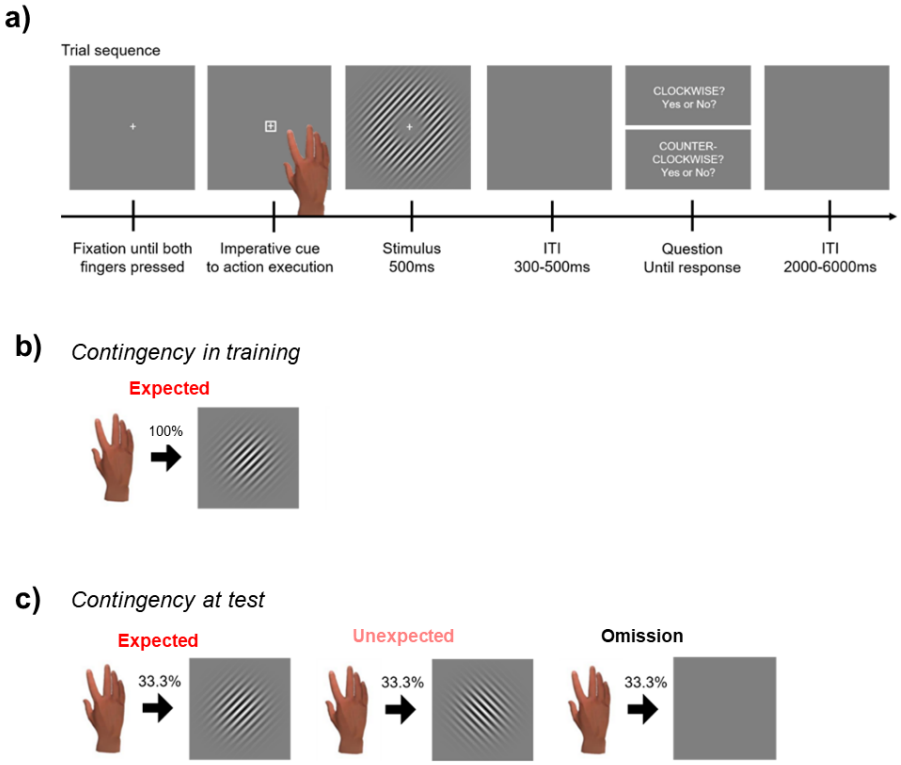

Figure 2: Illustration of experimental task. a) Participants performed a task where their actions produced oriented gratings which could be at expected or unexpected orientations, based on the training phase. We analysed brain activity evoked by the observed stimuli, and how this varied as a function of expectations. **b)** Prior to the scanning session, participants had formed probabilistic expectations, such that certain actions were always paired with certain gratings (i.e., 100% of trials were *expected*). **c)** However, during the task itself these *expected* stimuli were objectively unlikely: participants experienced the *expected* grating, the *unexpected* grating or an *omission* with equal probability (33.3%).

The experiment consisted of training and test sessions split over two consecutive days. On the first day, participants completed a training session outside of the scanner (600 trials, ~90 minutes). On these trials, participants performed the task while experiencing perfect statistical contingencies between executed actions and observed stimuli – e.g., index finger abduction always resulted in the presentation of a CCW oriented grating, while little finger abduction always resulted in a CW oriented grating. When arriving on the second day, participants completed a shorter but identical refresher session before entering the scanner room (120 trials, ~15 minutes). Cue-action and action-outcome mappings were counterbalanced across participants.

Participants then completed the main test session while in the MRI scanner, comprising 360 trials over ten scanning runs. The key change in this test session was that contingencies between action and outcome were disrupted. On *expected* trials, performing a given action yielded a grating stimulus congruent with the mapping learned in training (e.g., abducting the index finger generated a CCW grating - 33.3% of trials). However, there were an equal number of *unexpected* trials where the same action generated the opposite stimulus (e.g., abduct index, observe CW grating – 33.3% of trials). Moreover, on a further third of trials visual stimuli were completely omitted: participants performed the same action in response to the usual imperative shape cue, but after acting this cue disappeared, and no grating

stimulus was displayed (33.3% of trials). This patterning of trial types means that on a majority of trials, stimuli which were previously expected during training did not appear.

At the halfway point of both training and test sessions, the mapping between imperative shape cues (square and triangle) and to-be executed actions (index abduction or little abduction) was reversed, and participants were explicitly informed of the switch. This feature of the design removed any correlation between the shape cues and actual or expected grating orientations across the experiment.

We analysed BOLD activity evoked by grating stimuli (and their omission) to determine whether expectations shape visual brain activity even when previously acquired predictions are no longer reliable – that is, whether predictions are 'stubborn' (Yon, de Lange & Press, 2019).

fMRI acquisition and preprocessing. Images were acquired using a 3T Prisma MRI scanner (Siemens, Erlangen, Germany) using a 32-channel head coil. Functional images were acquired using an echo planar imaging (EPI) sequence (ascending slice acquisition, TR = 3.36 s, TE1/TE2 = 30/30.25 MS, 48 slices, voxel resolution: 3 mm isotropic). Structural images were acquired using a magnetisation-prepared rapid gradient-echo (MP-RAGE) sequence (voxel resolution: 1mm isotropic).

Images were preprocessed in SPM12. The first six volumes of each participant's data in each scanning run were discarded to allow for T1 equilibration. All functional images were spatially realigned to the first image and temporally realigned to the 24th (middle) slice. The participant's structural image was then coregistered to the mean functional scan. No smoothing was applied to functional images. All analyses of functional activity were conducted in the participants' native space (i.e., images were not transformed into a common space [e.g., MNI]), as the use of forward deformation fields involves implicitly smoothing the activity of multiple voxels, potentially limiting the sensitivity of multivoxel pattern analyses (though see Op de Beeck, 2010).

Regions of interest: Analyses were restricted to a region of interest (ROI) in the primary visual cortex (V1) of each participant, as this region contains neural populations sensitive to properties defining our visual stimuli (oriented gratings). An ROI in V1 was identified for each participant based on both anatomy and activity in two steps. In the first step, the boundaries of V1 were estimated using Freesurfer. This involved converting each participant's structural image into a cortical surface, and using the morphology of cortical folds to estimate the anatomical location of V1 (Hinds et al., 2008).

It is likely that each participant's visual cortex contains only a subset of neurons tuned to the specific stimuli used in our experiment - and including irrelevant voxels can have a deleterious impact on decoding performance ('the curse of dimensionality', Bellman, 1961; Haynes, 2015). To restrict our analyses to only those voxels involved in representing our stimuli, in a second step we therefore identified V1 voxels which contained information distinguishing the stimuli used in our task. This was achieved using an independent searchlight decoding analysis (3 voxel radius) to identify voxels that could discriminate clockwise from counter-clockwise grating stimuli (decoding accuracy greater than chance [50%]) collapsed across experimental conditions. All subsequent analyses were conducted on these informative V1 voxels.

Multivariate decoding analyses: Multivariate pattern analyses were implemented using the TDT toolbox (Hebart et al., 2015). In each analysis, a linear support vector machine (SVM) was trained to discriminate which visual grating (clockwise tilt or counter-clockwise tilt) was observed on a given trial from BOLD activity across voxels. The initial step in each analysis

was the specification of a general linear model (GLM) in SPM12, including a separate regressor for each stimulus type (e.g., a clockwise grating, a counter-clockwise grating, or a visual omission) in each of the expectation conditions (e.g., expected clockwise, expected counter-clockwise) in each scanning run. In conditions where grating stimuli were presented, regressors were modelled to the onset of the observed stimulus. On omission trials, the regressors were modelled to the same time point in the trial (i.e., when a stimulus would normally appear). Movement parameters were included as nuisance regressors, and all model regressors were convolved with the canonical haemodynamic response function. This GLM generated ten beta images (one for each scanning run) for each stimulus type (clockwise, counter-clockwise, omission) in each expectation condition that were used for subsequent decoding analyses.

To test for the presence of multivariate 'sharpening', beta images were grouped into conditions where visual stimuli were presented and stimuli were either *expected* (e.g., expect clockwise, observed clockwise) or *unexpected* (e.g., expect clockwise, observe counter-clockwise). This yielded 40 beta images – 20 where stimuli were *expected* (ten clockwise, ten counter-clockwise) and 20 where stimuli were *unexpected* (also ten of each stimulus). Separate SVMs were trained and tested on the 20 beta images in each expectation condition, using a leave-one-out cross-validation procedure. For each decoding step, 18 images from nine scanning runs were used to estimate a linear discriminant function separating clockwise and counter-clockwise gratings, which was then applied to the remaining two beta images to classify them as either 'clockwise' or 'counter-clockwise'. The procedure resulted in ten decoding steps, where each step reserved beta images from one of the ten scanning runs for classifier testing. The SVM's accuracy was calculated as the proportion of correctly classified images across all decoding steps. Accuracy was then compared between *expected* and *unexpected* conditions to determine whether information about observed stimuli varied as a function of learned expectation – where higher decoding accuracies suggest more informative underlying patterns, and a higher signal-to-noise ratio.

A separate multivariate decoding analysis was also conducted to determine whether information about the expected stimulus could be identified on trials where stimuli were omitted entirely (Aitken et al., 2020; Hindy et al., 2016; Kok et al., 2014; Smith & Muckli, 2010). To this end, an SVM was trained on the 20 beta images capturing patterns of activity on omission trials – where participants could either expect a clockwise (ten) or counter-clockwise (ten) stimulus based on prior learning. This SVM was trained to classify the *expected* stimulus rather than the actual stimulus – as no stimuli were presented on omission trials. The same run-wise cross-validation was used, with decoding accuracy reflecting the SVM's accuracy in classifying which stimulus the participant would be expecting on this trial – given the movement they have just performed. Above chance decoding accuracy in this analysis would indicate that V1 contains reliable information about momentary expectations – a possible signature of top-down predictions being supplied to early visual cortex from other parts of the brain.

Results

Expectations influence perceptual decisions

Reaction times and choice accuracies were compared on *expected* and *unexpected* trials during the test session to determine whether previously learned expectations influence perceptual performance. This was achieved via ANOVAs including factors of expectedness (*expected, unexpected)* and scanning run (*1-10*), allowing us to analyse effects of learned expectation and whether they change across the experiment as participants experience more and more events that undermine previously learned associations. Complementary

Bayes Factors (BFs) were calculated in JASP to clarify where nonsignificant results suggested support for the null hypothesis.

Our analysis of decision accuracy revealed that observers made more accurate perceptual choices on *expected* – M (SEM) = .98 (.005) - compared to *unexpected* trials - M (SEM) = .96 (.005); $F_{1,19}$ = 18.47, p<.001. This effect did not interact with scanning run – $F_{9,171}$= 1.65, p= .103, BF_{01}= 0.883 - suggesting this behavioural advantage did not differ between early and late runs of the experiment (see Figure 3).

Comparable analysis of median reaction times found that decisions were numerically faster on *expected* – M (SEM) = .66 secs (.013) - compared to *unexpected* trials – M (SEM) = .69 secs (.019) - though this difference was nonsignificant ($F_{1,19}$ = 3.16, p= .091, BF_{01}= .004). Moreover, this analysis also found no interaction between this effect and scanning run ($F_{9,171}$ = 1.577, p= .126, BF_{01}= 6.478) – a pattern which would have been expected if previously learned predictions influenced reaction times in early phases of the experiment, but these effects washed out in subsequent blocks (see Figure 3).

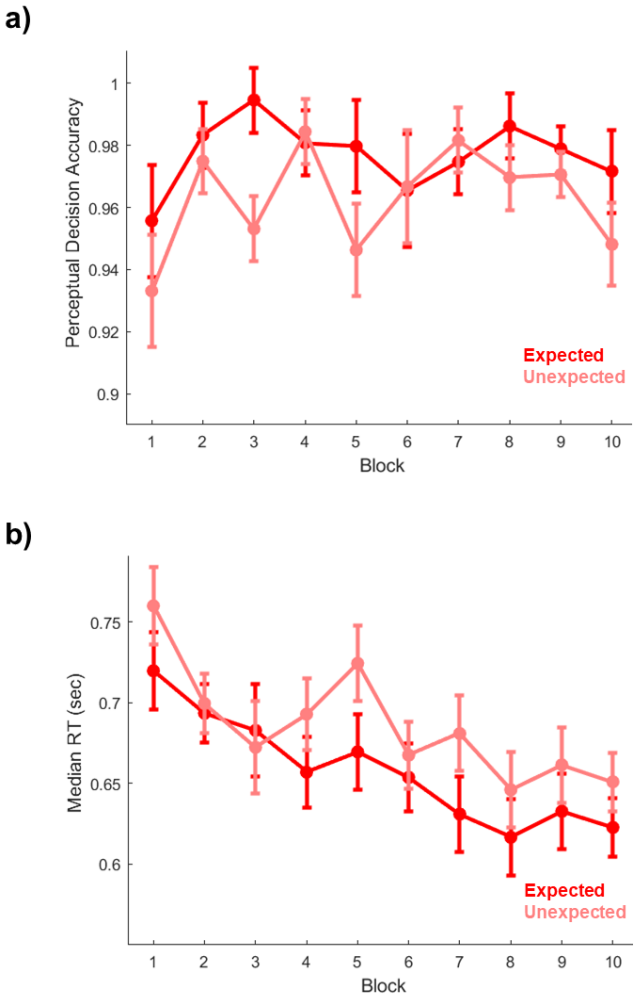

Figure 3: Expectations continue to shape behaviour in later experimental blocks. (**a**) Observers made more accurate perceptual decisions for expected compared to unexpected events, and this behavioural advantage did not change systematically across scanning blocks. (**b**) Observers had numerically faster reaction times on expected trials too, but this effect was non-significant. There was also no sign of this behavioural pattern changing over blocks. Both patterns suggest that expectations

established outside of the scanner continued to shape behaviour, even though expectations ceased to be reliable.

The effects of expectation on behaviour identified here suggest that prior learning continued to shape perceptual decision making, even though 'expected' events were improbable in the new test context. Moreover, the absence of any interaction between expectation and scanning run suggests that any effects of expectation on perceptual choice operated in a similar way across the scanning session – and did not wane as participants experienced more and more events that violated previous learning.

Expectations sharpen sensory representations in primary visual cortex

Support vector machines (SVMs) were used to decode from V1 the stimuli participants observed (clockwise, counter-clockwise), separately on *expected* and *unexpected* trials. Comparing decoding accuracies from primary visual cortex revealed superior decoding of expected – M(SEM) = 59.8% (1.6) – compared to unexpected stimuli – M(SEM) = 54% (1.6), t_{19} = 2.79, p = .012, dz = .624. This decoding benefit did not change across the experimental runs – $F_{9,171}$ = .699, p = .709, BF_{01} = 1829.47. This pattern of results is consistent with the idea that top-down predictions continue to alter activity in early visual areas even in environments where these expectations are no longer reliable.

However, comparing decoding effects for scanning run from these cross-validated SVMs is a conservative way of investigating whether effects change across the scanning session. This is because even when the classifier is *tested* on images from Block 10, it has been trained on images from Blocks 1-9. Thus, expectation effects present in later scanning blocks could be potentially contaminated by expectation effects in earlier blocks.

To further evaluate whether this expectation effect on decoding accuracy differed in early and late phases of the experiment, we therefore conducted additional analyses where classifiers were trained and tested exclusively on data from one half of the experiment. These SVMs were trained and tested using the same leave-one-out cross-validation procedure describe above, except our 'early runs' classifier was trained and tested on beta images from Runs 1-5 and our 'late runs' classifier on beta images from Runs 6-10. We then analysed decoding accuracies as a function of expectation (expected, unexpected) and experiment phase (early half, late half, see Figure 4d).

This analysis also revealed no interaction between expectation and experiment phase – $F_{1,19}$= .006, p = .940, BF_{01} = 8.346, suggesting that influences of expectation on neural decoding performance did not differ between early and late phases of the experiment. Such an interaction may have been expected if previously learned predictions exert a strong influence on neural representations in early phases of the experiment, with this influence waning in later phases as observers learn that past predictions no longer apply. Instead, these results are more consistent with the idea that predictions are 'stubborn' - continuing to operate despite evidence from the environment that old expectations are unlikely to come true.

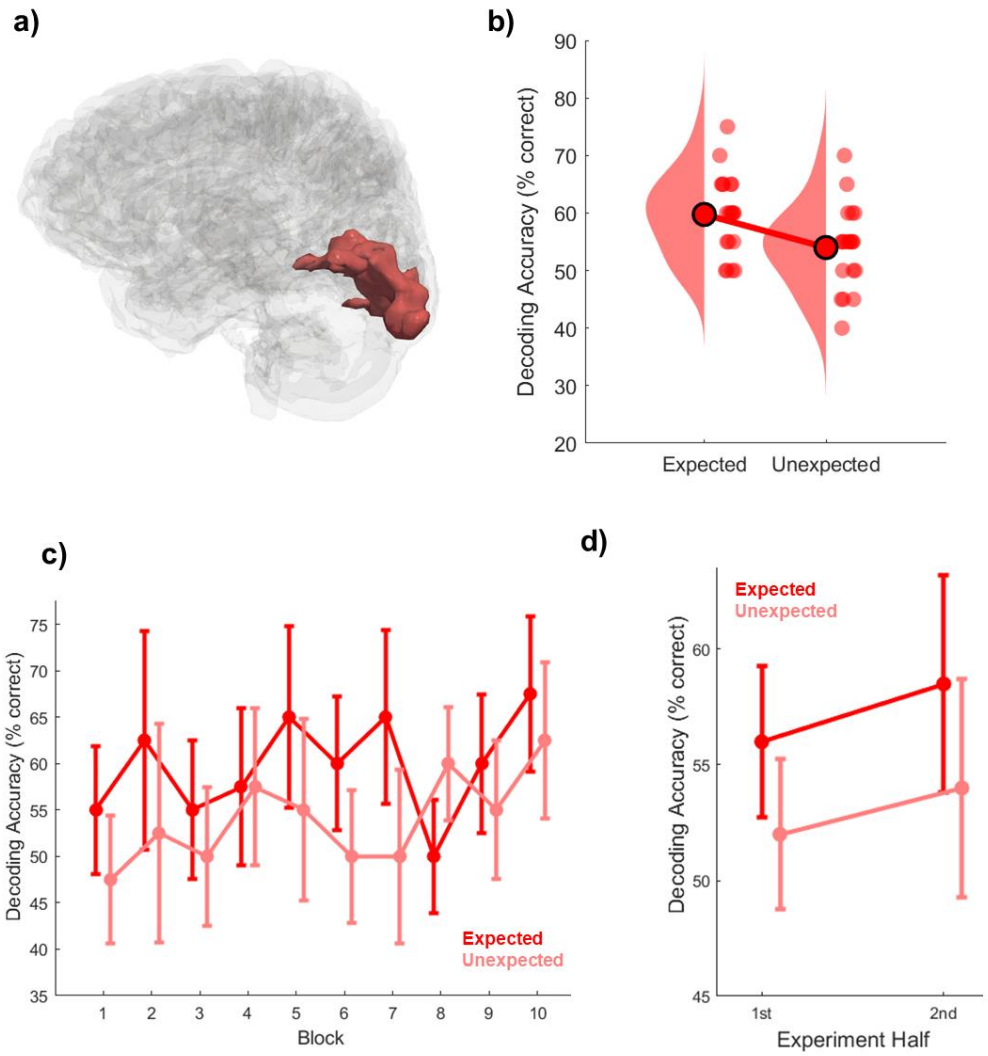

Figure 4: Enhanced decoding of expected events in primary visual cortex. (a) Illustration of primary visual cortex in native space of a single participant (**b**) Pattern classification accuracy across expectation conditions – revealing superior stimulus representations on expected trials. (**c**) When this decoding analysis is broken down by block, there is no discernible change in effects across scanning runs. (**d**) If decoding analyses are conducted separately on data from each experiment half, equivalent effects of expectation are found on decoding accuracies in early and late phases of the task.

Enhanced multivariate representations are related to changes in univariate activity patterns

These decoding analyses replicate previous reports of how expectations shape processing in visual cortex (Kok et al, 2012; Yon et al, 2018). Previously, we have found that improvements in decoding are associated with characteristic changes in univariate activity – such that expectations lead to a suppression of activity in voxels tuned *away from* the expected stimulus (Yon et al, 2018). Suppressing unexpected noise in this fashion provides

a plausible mechanism via which top-down predictions improve signal-to-noise across the sensory population, yielding better decoding when expectations are valid.

To investigate this possibility, here we conducted an analogous analysis to Yon et al (2018) - analysing how expectations change stimulus-specific patterns of BOLD activity. First, a t-test comparing activity evoked by clockwise and counter-clockwise gratings was used to classify the stimulus preference of each V1 voxel in a binary fashion (t>0 = clockwise preferred, t<0 = counter-clockwise preferred). This makes it possible to analyse stimulus-related activity for each voxel both as a function of expectations (e.g., was the stimulus expected or unexpected?) and preference (e.g., was the presented stimulus the one the voxel is most responsive to or not?). Analysing univariate BOLD activity (beta estimates) with a 2 x 2 expectation-by-preference ANOVA revealed no main effect of expectation – $F_{1,19} = .542$, p=.471 – and interestingly, also no interaction between expectation and stimulus preference in the univariate signal – $F_{1,19} = .108$, p=.746.

While this interaction was not significant, we nonetheless conducted further exploratory analyses to establish whether the decoding effects seen in this experiment were related to qualitatively similar changes in univariate activity patterns as seen in our previous work – specifically, a suppression of activity in units tuned away from the expected stimulus. To evaluate this possibility, we calculated two participant-wise effect scores. The first score was the effect of expectation on multivariate decoding (expected decoding accuracy – unexpected decoding accuracy) where positive numbers entail superior decoding on expected trials. The second score was the effect of expectation on univariate activity in voxels tuned away from the presented stimulus (expected non-preferred BOLD – unexpected non-preferred BOLD). Negative numbers on this effect score indicate a relative suppression of activity in voxels tuned to stimuli that are not currently expected. Correlating these two effects across participants revealed a significant negative relationship – $r_{20} = -.492$, p=.028 – such that those participants who showed the greatest multivariate decoding advantage for expected signals also showed the greatest univariate suppression in units tuned to unexpected events (see Fig 5). The same effect holds if calculating a non-parametric correlation – $T_{20} = -.369$, p = .032. This relationship also holds if we compute each participant's univariate interaction effect between expectation and stimulus preference – i.e., (expected preferred BOLD – unexpected preferred BOLD) - (expected non-preferred BOLD – unexpected non-preferred BOLD) - where higher numbers indicate relatively stronger suppression of activity in voxels tuned away from (rather than towards) expected stimuli. Correlating this interaction term with each participant's multivariate decoding effect also revealed a significant relationship - $r_{20} = .475$, p=.034. These patterns suggest that the underlying mechanisms at play are similar to those found in our previous work, even if the univariate patterns were not present at the group level.

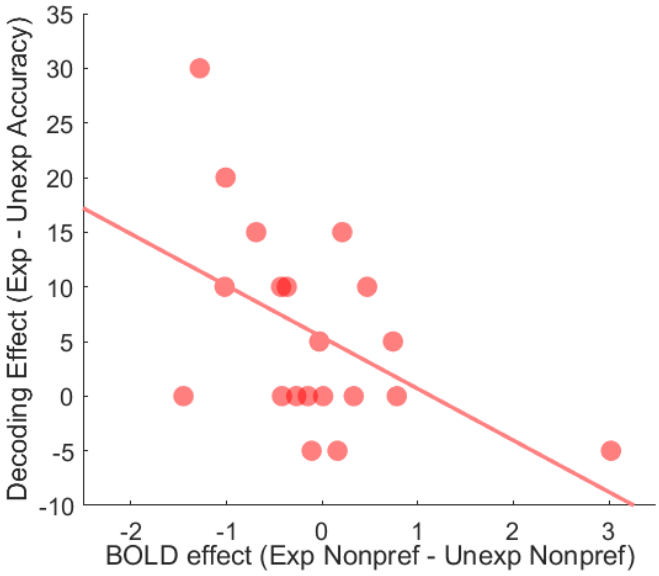

Figure 5: Superior decoding of expected events is related to suppression of activity in voxels tuned to the unexpected. The scatterplot displays the relationship between univariate and multivariate effects of expectation in V1. The y axis displays differences in multivariate decoding for expected and unexpected – where positive numbers indicate superior classification of expected stimuli. The x axis displays differences in univariate BOLD activity (beta values) between expected and unexpected trials in V1 voxels tuned *away from* the stimulus presented on that trial. Here, more negative numbers indicate a greater relative suppression of activity in voxels tuned to *unexpected* stimuli.

No reliable information about expectations when stimuli are omitted

Alongside stimulus-related activity, SVMs were also used to quantify information about the identity of stimuli that were expected but no signal was presented (omissions trials). If top-down predictions in V1 contain reliable information about expected stimuli, even when these are not presented, this analysis would reveal significantly above-chance decoding accuracy (Allefeld et al., 2016). However, one sample t-tests comparing decoding accuracy to chance (50%) revealed no reliable information about expected but omitted stimuli in V1 – $t_{19} = .311$, $p = .759$.

Discussion

Predictive models propose that we use prior experience to optimise perception of the here and now (Bar, 2004; de Lange et al., 2018). However, predictive perceptual systems in a changeable world face a challenge in determining how far perceptual inferences about the present should be guided by predictions from the past (O'Reilly, 2013; Yon, 2021). Models of learning suggest that this problem can be solved by estimating uncertainty in our predictions and our environments, but modern theories of prediction often incorporate the idea that expectations can be particularly 'stubborn' – brought to bear in new situations despite evidence that the world has changed (Yon et al., 2019).

Here we tested whether such 'stubborn predictions' are deployed in the visual brain, looking for signatures of predictive influence identified in previous work – but in novel conditions where the 'expected' is more likely *not* to be presented. Multivariate decoding analyses revealed that expectations enhanced the quality of sensory representations in primary visual cortex. This pattern is what is predicted by sharpening models of top-down expectation,

which assume that expectation signals increase the signal-to-noise ratio of sensory representations (de Lange et al, 2018; Press & Yon, 2019). At the group level, we did not find patterns in the univariate analyses like those we have identified in previous work, of relatively lower signal for expected events in voxels tuned to unexpected events (Kok et al., 2012; Yon et al., 2018). However, across participants we did see that enhancements of multivariate stimulus information on 'expected' trials were associated with a relative suppression of univariate signals in voxels tuned to unexpected events. Though more work is needed to establish the precise nature of the underlying mechanisms (e.g., probing multiple grating orientations to estimate population tuning curves), these findings are broadly in line with the findings from our earlier work.

The present study extends previous work by demonstrating that these expectation signals continue to reshape activity in these neural populations even when 'expected' events occur infrequently, and observers experience environmental statistics which strongly suggest that their old expectations should no longer apply. There was no sign of a decline in these effects as participants experienced more events inconsistent with previous learning.

One possible explanation for these findings is that human perceivers do not estimate the reliability of their predictions or the (in)stability of their environments in ways that normative learning models suggest that they should. For example, hierarchical learning models suppose that we can optimise learning about the relationships between events if we also estimate the rate at which our environment changes – learning more from new evidence when we detect that our environment is beginning to shift (Behrens et al., 2007; Yu & Dayan, 2005). If our participants approached their task as an optimal learner might, the presence of unexpected and omitted visual outcomes should induce changes in meta-level beliefs about the stability of the environment, driving new learning to update (and abolish) old, inappropriate predictions. Contrary to this possible norm, we find that perceivers continue to deploy top-down predictions even when environmental contingencies mean that 'expected' events do not occur most of the time. An undue reliance on old predictions could arise if agents estimate the stability of their environment (and the reliability of their predictions) incorrectly – believing that the world is more stable than it really is, and thus that experiences from the past are more relevant to interpreting the present than they really are.

Exaggerated beliefs in the stability of our environments could be a general suboptimality in prediction and learning mechanisms that has been hitherto underappreciated. This could stem from general computational limitations in bounded agents like us: imprecise learning about the reliability of our predictions could stem from the fact that meta-level quantities (like volatility) are naturally difficult to estimate, because tracking the stability of our environment over time requires integrating over many more datapoints than tracking what's happening in the here and now (Yon & Frith, 2021). Given this computational constraint, it seems possible that observers form 'stubborn predictions' because their initial belief about the stability of the environment (established through perfectly contingent training) is not updated sufficiently quickly to drive (un)learning about perceptual predictions.

However, another possibility is that agents do not generally struggle to estimate environmental volatility, but are prone to specific biases when estimating the predictive power of their *actions*. Recent studies in reinforcement learning have suggested that when agents experience identical schedules of variable wins and losses, they are biased to perceive environments as more stable when rewards are predicted by their actions rather than equally predictive sensory cues (Weiss et al., 2021). Such effects could reflect a high-level belief about volatility and action – meaning that we expect the world to stabilise when we interact with it. Previously, we have argued that expectation mechanisms which use

sensory context or executed actions to predict upcoming events operate in functionally equivalent ways (Press et al., 2020) – and comparable influences of both kinds of prediction are observed in brain and behaviour. However, it remains possible that potential differences in volatility estimation between active and passive prediction are important for understanding how different kinds of expectation guide perceptual inference – since it may generally be true that mappings between action and outcome tend to be more reliable than mappings between different sensory cues. Armed with these background beliefs, agents may frequently mistake spurious associations between action and outcome for genuine contingencies (Yon et al., 2020).

These neuroimaging findings provide an interesting complement to recent work using similar techniques to study behavioural and computational consequences of prediction on perception as predictive relationships degrade (Thomas et al., 2022). This work reveals that when perfect predictive associations are replaced with weaker associations, prior learning can continue to bias perceptual decisions. Moreover, computational modelling suggests that such biases arise via altering both the starting points and rates of sensory evidence accumulation (Ratcliff et al., 2016; Yon et al., 2021; see also Wyart et al., 2012). These effects are compatible with the kind of neural mechanism we observe here – where activity in the sensory brain is weighted towards predicted outcomes, yielding 'sharper' representations of expected events.

There is however one possible discrepancy between this behavioural work and our present neuroimaging results. Here, we find evidence that perceptual predictions are retained in environments where previously expected events become objectively improbable (33% of trials), while Thomas et al. (2022) find that effects of expectation are completely abolished when predictive contingencies disappear (i.e. when 'expected' and 'unexpected' events are equiprobable in a subsequent test session [50%]). A possible reason for this difference is that participants in the present study received more training than those in Thomas et al. (2022), as well as being presented with a longer 'refresher' of the learned contingencies before disruption. Differences in this kind of experience may alter inferences agents make about the stability of their environments, and the stubbornness of resulting predictions. It is therefore important for future work to establish the boundary conditions of these phenomena: identifying when and why predictions are rapidly updated, and when they are stubbornly retained.

While we found evidence that stubborn predictions shape visual representations of expected and unexpected events, we did not find evidence for expectation signals in V1 when events were omitted. Such effects have been observed in multiple previous studies (e.g., Aitken et al., 2020; Hindy et al., 2016; Kok et al., 2014) when expectations are probed while probabilistic knowledge remains valid. Thus, it is possible that we did not observe any reliable expectation-related activity on omission trials in our task because predictive mappings were not sufficiently strong. However, an alternate possibility is that multivariate decoding is not the best tool to identify such signals. Other attempts to reveal 'templates' of expected but omitted stimuli have instead quantified the univariate activity of units that prefer the expected stimulus – based on independent data to establish voxel stimulus preferences separate from expectations. We did not include similar independent 'stimulus only' runs in the present experiment, and this may be a promising avenue for future work.

In conclusion, we have found that stimulus representations in primary visual cortex continue to be 'sharpened' in line with top-down expectations, even when the environment signals that 'expected' events are no longer likely. Evidence that our perceptual systems form such 'stubborn predictions' reveals important ways that learning and prediction in the human brain

may depart from principles laid down in optimal models – possibly due to intrinsic constraints on our ability to estimate different kinds of uncertainty. Understanding these bounds on human information processing will be critical for understanding the symbiotic relationship between learning and perception, and for determining when our predictive models improve and impair our ability to make sense of the world around us.

References

Aitken, F., Menelaou, G., Warrington, O., Koolschijn, R. S., Corbin, N., Callaghan, M. F., & Kok, P. (2020). Prior expectations evoke stimulus-specific activity in the deep layers of the primary visual cortex. *PLOS Biology*, *18*(12), e3001023. https://doi.org/10.1371/journal.pbio.3001023

Allefeld, C., Görgen, K., & Haynes, J.-D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *NeuroImage*, *141*, 378–392. https://doi.org/10.1016/j.neuroimage.2016.07.040

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Bellman, R. E. (1961). Adaptive Control Processes: A Guided Tour. In *Adaptive Control Processes*. Princeton University Press. https://doi.org/10.1515/9781400874668

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Friston, K. (2018). Does predictive coding have a future? *Nature Neuroscience*, *21*(8), 1019–1021. https://doi.org/10.1038/s41593-018-0200-7

González-García, C., & He, B. J. (2021). A Gradient of Sharpening Effects by Perceptual Prior across the Human Cortical Hierarchy. *Journal of Neuroscience*, *41*(1), 167–178. https://doi.org/10.1523/JNEUROSCI.2023-20.2020

Hansen, K.A., Hillenbrand, S.F., & Ungerleider, L. G. (2011). Persistency of priors-induced bias in decision behavior and the fMRI signal. *Frontiers in Neuroscience, 5,* https://doi.org/10.3389/fnins.2011.00029.

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

Hebart, M. N., Görgen, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*. https://www.frontiersin.org/article/10.3389/fninf.2014.00088

Heilbron, M., & Chait, M. (2018). Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? *Neuroscience*, *389*, 54–73. https://doi.org/10.1016/j.neuroscience.2017.07.061

Hinds, O. P., Rajendran, N., Polimeni, J. R., Augustinack, J. C., Wiggins, G., Wald, L. L., Diana Rosas, H., Potthast, A., Schwartz, E. L., & Fischl, B. (2008). Accurate prediction of V1 location from cortical folds in a surface coordinate system. *NeuroImage*, *39*(4), 1585–1599. https://doi.org/10.1016/j.neuroimage.2007.10.033

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience*, *19*(5), 665–667. https://doi.org/10.1038/nn.4284

Hogendoorn, H. (2022). Perception in real-time: Predicting the present, reconstructing the past. *Trends in Cognitive Sciences*, *26*(2), 128–141. https://doi.org/10.1016/j.tics.2021.11.003

Kok, P., Failing, M. F., & de Lange, F. P. (2014). Prior Expectations Evoke Stimulus Templates in the Primary Visual Cortex. *Journal of Cognitive Neuroscience*, *26*(7), 1546–1554. https://doi.org/10.1162/jocn_a_00562

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*(2), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034

Op de Beeck, H. P. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, *49*(3), 1943–1948. https://doi.org/10.1016/j.neuroimage.2009.02.047

O'Reilly, J. (2013). Making predictions in a changing world—Inference, uncertainty, and learning. *Frontiers in Neuroscience*, *7*. https://www.frontiersin.org/article/10.3389/fnins.2013.00105
Press, C., Kok, P., & Yon, D. (2020). The Perceptual Prediction Paradox. *Trends in Cognitive Sciences*, *24*(1), 13–24. https://doi.org/10.1016/j.tics.2019.11.003
Press, C., & Yon, D. (2019). Perceptual Prediction: Rapidly Making Sense of a Noisy World. *Current Biology*, *29*(15), R751–R753. https://doi.org/10.1016/j.cub.2019.06.054
Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007
Smith, F. W., & Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(46), 20099–20103. https://doi.org/10.1073/pnas.1000233107
Thomas, E., Rittershofer, K., & Press, C. (2022). *Updating perceptual expectations as certainty diminishes*. PsyArXiv. https://doi.org/10.31234/osf.io/z6xnd
Weiss, A., Chambon, V., Lee, J. K., Drugowitsch, J., & Wyart, V. (2021). Interacting with volatile environments stabilizes hidden-state inference and its brain signatures. *Nature Communications*, *12*(1), 2228. https://doi.org/10.1038/s41467-021-22396-6
Wyart, V., Nobre, A. C., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, *109*(9), 3593–3598. https://doi.org/10.1073/pnas.1120118109
Yon, D. (2021). Prediction and Learning: Understanding Uncertainty. *Current Biology*, *31*(1), R23–R25. https://doi.org/10.1016/j.cub.2020.10.052
Yon, D., Bunce, C., & Press, C. (2020). Illusions of control without delusions of grandeur. *Cognition*, *205*, 104429. https://doi.org/10.1016/j.cognition.2020.104429
Yon, D., de Lange, F. P., & Press, C. (2019). The Predictive Brain as a Stubborn Scientist. *Trends in Cognitive Sciences*, *23*(1), 6–8. https://doi.org/10.1016/j.tics.2018.10.003
Yon, D., & Frith, C. D. (2021). Precision and the Bayesian brain. *Current Biology*, *31*(17), R1026–R1032. https://doi.org/10.1016/j.cub.2021.07.044
Yon, D., Gilbert, S. J., de Lange, F. P., & Press, C. (2018). Action sharpens sensory representations of expected outcomes. *Nature Communications*, *9*(1), 4288. https://doi.org/10.1038/s41467-018-06752-7
Yon, D., Zainzinger, V., de Lange, F. P., Eimer, M., & Press, C. (2021). Action biases perceptual decisions toward expected outcomes. *Journal of Experimental Psychology. General*, *150*(6), 1225–1236. https://doi.org/10.1037/xge0000826
Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026
Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308. https://doi.org/10.1016/j.tics.2006.05.002

Acknowledgements

This work was supported by the Leverhulme Trust. CP, FPdL and PK acknowledge support from the European Research Council.

Data availability

Data is available at https://osf.io/7svmg/.

Author contributions

Daniel Yon – Conceptualisation, Data Curation, Formal Analysis, Investigation, Methodology, Visualisation, Writing (Original Draft, Review & Editing)

Emily Thomas – Conceptualisation, Data Curation, Investigation, Methodology, Writing (Review & Editing)

Sam Gilbert – Conceptualisation, Writing (Review & Editing)

Floris de Lange – Conceptualisation, Writing (Review & Editing)

Peter Kok – Conceptualisation, Writing (Review & Editing)

Clare Press – Conceptualisation, Funding Acquisition, Supervision, Writing (Review & Editing)

Rights retention