Accepted at *Cognition*, 17th October 2022

Updating perceptual expectations as certainty diminishes

Emily R. Thomas[1,2*], Kirsten Rittershofer[1*♯], & Clare Press[1,3]

1. Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK.
2. Neuroscience Institute, New York University School of Medicine, 550 1st Ave, New York, NY10016, US.
3. Wellcome Centre for Human Neuroimaging, UCL, 12 Queen Square, London, WC1N 3AR, UK.

* Equal contribution

♯ k.rittershofer@bbk.ac.uk

Abstract

Forming expectations about what we are likely to perceive often facilitates perception. We forge such expectations on the basis of strong statistical relationships between events in our environment. However, due to our ever-changing world these relationships often subsequently degrade or even disappear, yet it is unclear how these altered statistics influence perceptual expectations. We examined this question across two studies by training participants in perfect relationships between actions (index or little finger abductions) and outcomes (clockwise or counter-clockwise gratings), before degrading the predictive relationship in a test phase – such that 'expected' events followed actions on 50-75% of trials and 'unexpected' events ensued on the remainder. Perceptual decisions about outcomes were faster and less error prone on expected than unexpected trials when predictive relationships remained high and reduced as the relationship diminished. Drift diffusion modelling indicated that these effects are explained by shifting the starting point in the evidence accumulation process as well as biasing the rate of evidence accumulation – with the former reflecting biases from statistics within the training session and the latter those of the test session. These findings demonstrate how perceptual expectations are updated as statistical certainty diminishes, with interacting influences speculatively dependent upon learning consolidation. We discuss how underlying mechanisms optimise the interaction between learning and perception – allowing our experiences to reflect a nuanced, ever-changing environment.

Keywords: prediction; expectation; uncertainty; perception; DDM

1. Introduction

It is essential that we learn about relationships between events in our world, such that we can perceive and interact with it appropriately. For example, to survive we must learn that stamping on the brake pedal makes the car stop, or that when we send motor commands to grasp a mug of tea we see fingers converging towards the target mug. We learn these relationships when there is a correlation between the activation of the event representations (Rescorla & Wagner, 1972). In associative terms, this contingency strengthens the association between representations, while in Bayesian terminology one event provides a 'prior' for the other, allowing us to predict it (Oaksford & Chater, 2007).

While strong statistical relationships help us to form expectations about upcoming perceptual events, it is common that the strength of such relationships subsequently degrades over time. For example, we foveate our hand frequently as an infant when we learn to grasp, but by adulthood we foveate our hand less frequently – instead often foveating the target object or something else entirely (Johansson et al., 2001). The sight of fingers moving together is still more common than any other singular visual input when we grasp, meaning that there is still a predictive relationship (Dickinson & Charnock, 1985), but it is now degraded. In extreme cases, a once strong relationship can even disappear entirely, for instance if we cease to receive tactile stimulation on a moving effector due to nerve damage.

It is however unclear how such statistical degradation influences perceptual expectations. Perhaps counterintuitively, findings from visual and tactile cognition demonstrate a range of persisting predictive influences on perceptual processing that no longer reflect the statistics of the environment (Press, Thomas, et al., 2020; Yon et al., 2018, 2021; Yon & Press, 2017, 2018), suggesting that predictions may be especially stubborn in the face of degraded relationships. For example, when participants are presented with no relationship between actions and outcomes, predictions that have been previously formed based upon a perfectly correlated training phase (Press, Thomas, et al., 2020) or through a lifetime of experience (Yon et al., 2018, 2021; Yon & Press, 2017, 2018) continue to influence perceptual processing.

It is suggested that priors are ‘stubborn’ (Yon et al., 2019), or difficult to update, and are thought by some to explain curious effects of perception in synaesthesia – e.g., numbers appearing coloured due to statistical regularities encountered during childhood (Yon & Press, 2014). Although there are a number of conceptual differences, work from the animal learning literature also indicates that predictions may not be updated at all as correlations degrade. For example in a classic study, Rescorla (1992, see also 1993) trained rats to learn action-outcome associations e.g., lever press-sucrose, and recorded lever pressing behaviour after devaluing the outcome by pairing it with nausea-generating Lithium Chloride. Such lever pressing was identical regardless of whether rats experienced intervening extinction phases (where actions no longer produce the outcome). These findings were taken to suggest that action-outcome associations remain unaltered through extinction phases (Crimmins et al., 2021), and that such resistance to change may reflect a general principle of learning – applying also to stimulus-outcome learning (Delamater, 2012).

However, if predictions really exhibit such stubbornness, this would be Bayes-suboptimal. Bayesian frameworks assume that an ideal observer generates a precision-weighted combination of expectations (prior) and inputs (likelihood) to determine what they perceive (posterior; Ernst & Banks, 2002; Garrido et al., 2013; Skora et al., 2021; Yon et al., 2021). If the precision on the prior is high – due to strong statistical relationships within our environment – perception will be heavily weighted towards expectations, in turn rendering us less sensitive to inputs that could update the priors (Press, Kok, et al., 2020a). Nevertheless, the precision on the input would also be high in most of these studies indicating stubborn predictions, since the input is suprathreshold and unambiguous. One must also therefore consider that stubbornness of expectations could be generated due to estimating our environment to be more stable than it is (Behrens et al., 2007). Specifically, in stable environments – where relationships between events remain consistent across time – new events will not provide much information about the structure of the world and our learning rate is assumed to be low. The manipulated statistical environment in the studies discussed is indeed relatively stable

before the test phases – considering that there is a perfect relationship between actions and outcomes in training phases, single events would provide little information about the structure of the environment. When the environment changes we may therefore not update our predictions accordingly because we have erroneously estimated these relationships to be stable. Such errors may be widespread because it is notoriously difficult for us to estimate uncertainty within our environment (Yon & Frith, 2021).

The present studies were therefore designed to address whether and how reducing the strength of statistical relationships diminishes predictive influences on perceptual decisions. It has typically been found that expected events are perceived with greater speed and accuracy than unexpected events (Bar, 2004; Kok et al., 2017; Ouden et al., 2010; Palmer, 1975; Yon et al., 2021; Yon & Press, 2018) and it is assumed that a stronger (more precise) expectation would generate a greater difference between expected and unexpected trials. We wished to determine whether predictions forged on the basis of strong relationships are really so resistant to change or whether differing levels of degradation would be reflected accordingly in the perceptual decisions. If such expectations are updated we also wished to understand how.

In two experiments, participants were trained with perfect relationships between actions (index or little finger abductions) and outcomes (clockwise or counter-clockwise oriented gratings). These predictive relationships were subsequently disrupted such that the trained outcome followed action on some of the trials (expected trials) and the other grating ensued on the remaining (unexpected) trials. Participants answered whether a particular grating orientation was presented, and we analysed response times and error rates. A pilot experiment confirmed that influences of expectation on perceptual decisions could be observed in such a setting (see Supplementary Information). Experiment 1 compared perceptual decisions in groups who experienced differing levels of disruption (75%, 67% or 50% expected trials). Experiment 2 replicated findings in Experiment 1, while isolating influences of learning within training and

test phases. We also examined start and drift biasing parameters of the perceptual decision process (drift diffusion modelling) to shed light on the mechanisms underlying degradation.

2. Experiment 1

Experiment 1 trained three groups of participants in perfect mappings between actions (index and little finger abductions) and outcomes (clockwise [CW] and counter-clockwise [CCW] gratings). Participants answered “yes/no” to one of two response questions probing “CW grating presented?” or “CCW grating presented?”. The question changed frequently to ensure that participants learned action-outcome associations rather than action-response associations. To test the influence of statistical degradation on reaction times and error rates, the three groups experienced different levels of degradation to the trained mappings during a subsequent test phase. Participants in the group with the lowest degradation were presented with expected action outcomes on 75% of trials and unexpected outcomes on the remaining 25%. We compared the effects from this group against a group who were presented with higher degradation (67% group, whereby the expected events are presented on 67% of trials and the unexpected events on 33%) and against a group where there was no longer a predictive relationship at test (50% group, whereby expected and unexpected outcomes are each presented on 50% of trials).

2.1. Methods

Participants

Twenty-four participants were recruited in each of the 50% (mean age = 27.17 years, SD = 7.03), 67% (mean age = 29.67 years, SD = 7.07) and 75% (mean age = 25.92 years, SD = 5.75) degradation groups, totalling 72 participants overall (power determined via piloting; see Supplementary Information). Participants were recruited via Prolific, completed the experiment online, and were paid a small honorarium for participation. Participants whose accuracy on either trial type (expected or unexpected) was not above chance (50%), or who had fewer than 300 trials remaining after removing those with very short or long response times (RTs; see

"Analysis"), were excluded from the analysis. Therefore, one participant was a replacement for a participant whose accuracy was below chance on unexpected trials. All participants reported normal or corrected to normal vision.

Stimuli

Sinusoidal grating stimuli were created using MATLAB and presented against a grey background. A Gaussian filter enveloped the grating stimuli to create Gabor patches of 80% Michelson contrast. The CW grating was oriented at 45° (relative to the hypothetical vertical mid-point i.e., 12 o'clock, 90°) and the CCW grating at 135°.

Procedure

The experiment was programmed to run using Gorilla (www.gorilla.sc) for online experiments, and participants took part on either a laptop or desktop computer. Participants were instructed to set maximum screen brightness in an attempt to reduce variability in viewing conditions. During the 100% training phase, participants abducted either the index or little finger of their right hand that perfectly predicted the orientation (CW or CCW) of the stimulus. Participants were instructed to hold down the 'c' and 'm' keys with their respective right index and little fingers on the computer keyboard until an imperative cue (square or triangle) indicated the finger to abduct. As soon as the cued finger was abducted (i.e., the corresponding key was released), the visual stimulus was presented for 500 ms in an annulus that surrounded the response question in shorthand ('CW?' or 'CCW?'; see Fig. 1). This response text appeared in synchrony with the onset of the grating stimulus and remained on screen until a response was selected. The stimulus orientation probed by the question alternated every mini-block of 36 trials. Responses were made using the left thumb on the 'a' and 'z' keys, for 'yes' and 'no', respectively. Following the response, a variable inter-trial interval of 1500-3000 ms was presented before the start of the next trial.

In the test phase 24 hours later, participants were allocated to one of three groups which only differed by level of degradation of the learned action-outcome relationship from the 100%

training phase – where the expected orientation was either presented on 75%, 67%, or 50% of trials. The unexpected, and therefore orthogonal, orientation was presented on the remaining trials in each condition. There were 360 main experimental trials in each session – the anticipated maximum number of trials that participants could undertake in a single session before exhibiting fatigue effects. Participants completed 32 practice trials before proceeding to the main trials in the training phase. Trial order was randomised and the specific action-stimulus relationship was counterbalanced across participants. The imperative cue indicating the required action was randomised across participants and this cue-action mapping was reversed at the midpoint of each session.

Analysis

All trials with RTs below 100 ms or larger than 2500 ms were excluded. For each participant, the effect of expectation on RTs for correct trials and proportion of errors (PEs) was calculated by subtracting the RT/PE for expected trials from unexpected trials. Positive

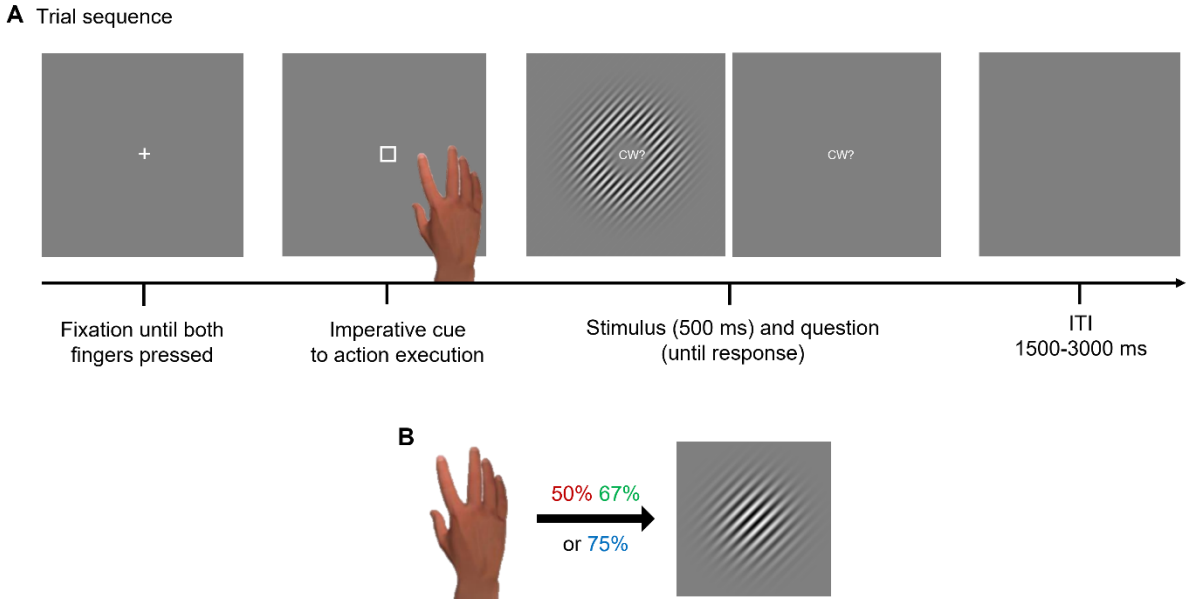

Figure 1. Experiment 1 task. (A) A centrally presented visual cue instructed participants to abduct either their index or little finger. In the 100% training phase, each finger abduction perfectly predicted an oriented stimulus and participants were required to respond (yes/no) to the presented question – indicating whether the stimulus was clockwise ('CW?') or counter-clockwise ('CCW?'). **(B)** In the subsequent degraded test phase, participants performed the same task but were allocated into one of three groups where the presented orientation was consistent with the previously learned mapping on either 75%, 67%, or 50% of trials.

values reflect faster and less error prone responses on expected compared to unexpected trials.

2.2. Results and Discussion

A one-way ANOVA on the RT expectation effect showed a significant effect of Degradation ($F(2,69) = 7.35$, $p = .001$, $\eta_p^2 = .18$; Fig. 2A). Participants in the 75% group showed a larger RT expectation effect than both the 67% ($t(46) = 2.55$, $p = .014$, $d = 0.74$) and 50% groups ($t(46) = 3.76$, $p < .001$, $d = 1.09$), while the RT effects in the 67% and 50% groups did not differ ($t(46) = 1.18$, $p = .246$, $d = 0.34$). One-sample t-tests revealed significant RT expectation effects for the 75% ($M = 44.14$ ms, $SD = 33.63$, $t(23) = 6.43$, $p < .001$, $d = 1.31$) and 67% ($M = 19.47$ ms, $SD = 33.46$, $t(23) = 2.85$, $p = .009$, $d = 0.58$) groups, but not for the 50% group ($M = 8.31$ ms, $SD = 32.33$, $t(23) = 1.26$, $p = .221$, $d = 0.26$).

A similar pattern was observed for the effect of Degradation on PEs ($F(2,69) = 9.87$, $p < .001$, $\eta_p^2 = .22$; Fig. 2B). Participants in both the 75% ($t(46) = 3.90$, $p < .001$, $d = 1.13$) and 67% groups ($t(46) = 3.42$, $p = .001$, $d = 0.99$) showed greater PE expectation effects than the 50% group, while there was no difference between the effects in the 67% and 75% groups ($t(46) = 2.00$, $p = .052$, $d = 0.58$). Both the 75% ($M = 0.05$, $SD = 0.06$, $t(23) = 4.40$, $p < .001$, $d = 0.90$) and 67% ($M = 0.03$, $SD = 0.02$, $t(23) = 5.22$, $p < .001$, $d = 1.07$) groups showed significant expectation effects, but the 50% group did not ($M = 0.002$, $SD = 0.02$, $t(23) = 0.37$, $p = .714$, $d = 0.08$; see Supplementary Information for consideration of the development of these effects across trial).

These results demonstrate that participants are faster, and more accurate, at making perceptual decisions about expected, relative to unexpected, events when the predictive relationship between action and outcome remained highest at test (75%). Furthermore, expectation effects became smaller when the relationship was more degraded (67%), and were entirely absent when the relationship was eliminated (50%), suggesting that expectations are updated when the statistics of the environment are altered.

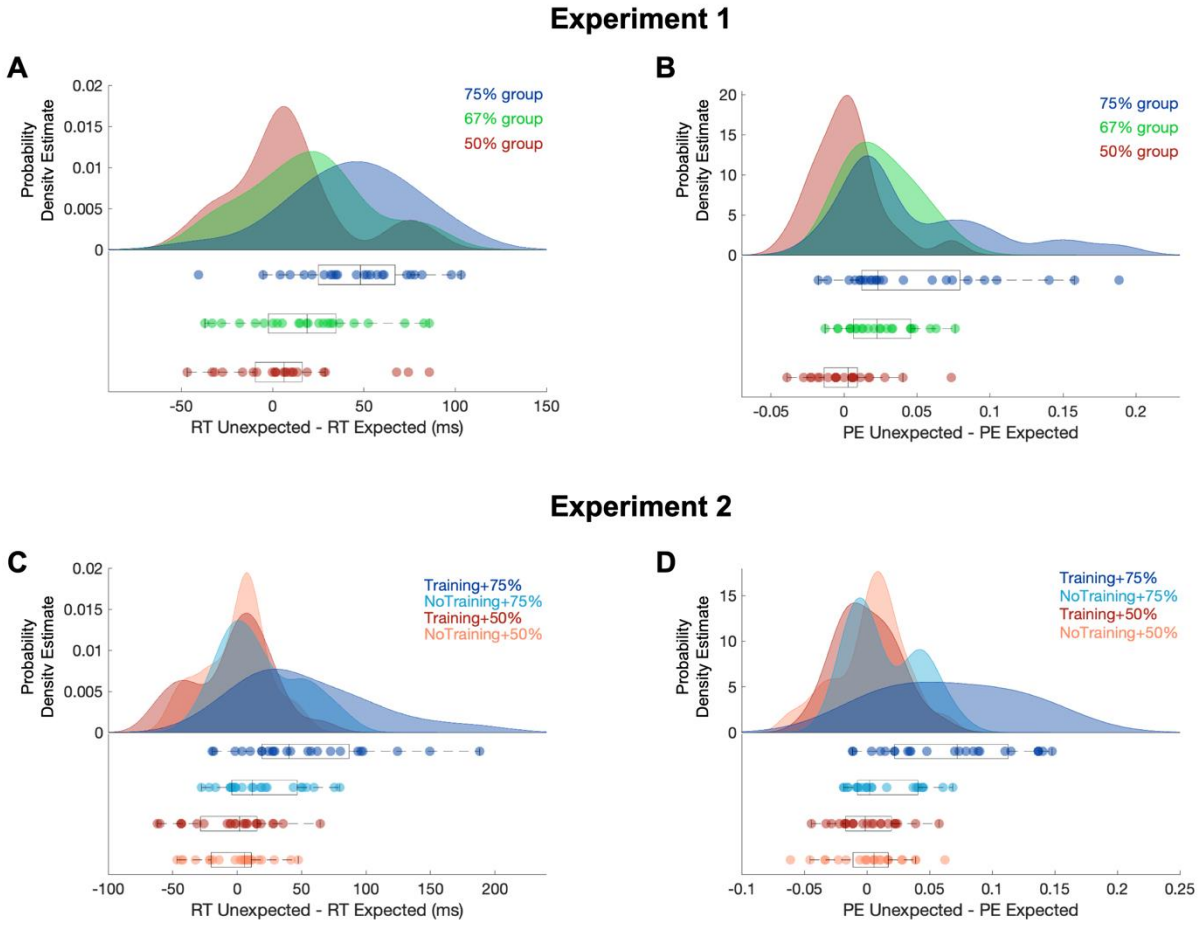

Figure 2. Experiment 1 and 2 Results. (A) Experiment 1 RT expectation effect for the 50%, 67%, and 75% groups (*N* = 24 per group). Participants in the 75% group showed a larger RT expectation effect than the 67% and 50% groups. **(B)** Experiment 1 PE expectation effect for the three groups. Participants in the 75% and 67% groups both showed larger expectation effects than the 50% group. **(C)** Experiment 2 RT expectation effect for the 75% and 50% Training and NoTraining groups (*N* = 24 per group). Significant expectation effects were present in both 75% groups, while these were larger in those who had experienced a preceding 100% training phase. No effects were observed in either 50% groups. **(D)** Experiment 2 PE expectation effect for the four groups. The patterns were identical to those in RTs (C). The results are plotted with raincloud plots (Allen et al., 2019) displaying probability density estimates (upper) and box and scatter plots (lower). Boxes denote lower, middle and upper quartiles, whiskers denote 1.5 interquartile range, and dots denote difference scores for each participant.

3. Experiment 2

Experiment 1 thereby demonstrated that participants were fastest and least error prone when responding to expected relative to unexpected events in situations where the statistical action-outcome relationship had the smallest amount of degradation at test (75% group). Expectation effects reduced when these relationships were further degraded (67% group) and disappeared

when actions were no longer predictive of the outcome. Such differences according to the level of degradation demonstrate that we likely do not stubbornly, and suboptimally, maintain perceptual predictions established under strong statistical relationships when the strength of these relationships alters – but rather are able to update our expectations to match the new statistics of our environment and inform perceptual decision making.

Experiment 2 was designed to serve two purposes. Firstly, it aimed to replicate the core finding from Experiment 1, and secondly, it compared the magnitude of these effects against participant groups who did not experience an initial 100% training phase. The latter manipulation was added to confirm that the effects observed in Experiment 1 result from updating expectations as relationships degrade. That is, in principle the same effects could be observed if learning during the training phase is discarded at test, and therefore only reflect differences in expectations during the test phase – regardless of any previous learning. Thus, if the effects observed in Experiment 1 indeed result from updating of previously learned expectations with changing statistical relationships, we would not expect the same effects in these groups where the training phase is omitted. However, if the effects purely represent the statistics of the environment during the test session regardless of any previous learning, we would expect effects to be similar across groups with and without a preceding training phase.

3.1. Methods

Participants

Twenty-four participants were recruited in each of the Training+50% (mean age = 29.83 years, *SD* = 10.48), NoTraining+50% (mean age = 30.79 years, *SD* = 11.12), Training+75% (mean age = 24.83 years, *SD* = 4.98) and NoTraining+75% groups (mean age = 32.63 years, *SD* = 13.13), totalling 96 participants overall. Participants were recruited via Prolific, completed the experiment online and were paid a small honorarium for participation. Three participants were replacements for those who had fewer than 300 trials after removing those with very short or long RTs.

Procedure

The procedure for the 75% and 50% Training groups in Experiment 2 was similar to the 75% and 50% groups in Experiment 1. The only difference was that participants also completed a short training refresher at the beginning of day 2 – consisting of 36 trials with 100% action-outcome relationships. We also collected data from two groups without a training phase, i.e., 75% and 50% NoTraining groups. For participants in the NoTraining groups, the test session was identical to that in the Training groups, and they also completed the same overall number of trials in a previous session. Crucially though, for the NoTraining groups, participants did not receive training in action-stimulus mappings in the first session, but instead were presented with the same grating stimuli without a requirement for performing any actions. Stimulus onset was instead triggered automatically in this session, 1500 ms after the fixation cross at the start of the trial. This way, participants in all groups had the same exposure to the task sessions and stimuli, but were not able to learn a predictive relationship between actions and grating orientations[1]. All other parameters of the task were the same across groups.

3.2. Results and Discussion

A two-way ANOVA on the RT expectation effect revealed a significant main effect of Degradation ($F(1,92) = 26.05$, $p < .001$, $\eta_p^2 = .22$) and Training ($F(1,92) = 4.84$, $p = .030$, $\eta_p^2 = .05$), as well as a significant interaction between these two factors ($F(1,92) = 6.60$, $p = .012$, $\eta_p^2 = .07$; Fig. 2C). Both 75% groups showed significant RT expectation effects (Training+75%: $M = 53.27$ ms, $SD = 52.06$, $t(23) = 5.01$, $p < .001$, $d = 1.02$; NoTraining+75%: $M = 18.01$ ms, $SD = 30.41$, $t(23) = 2.90$, $p = .008$, $d = 0.59$) as revealed by one-sample t-tests. However, pairwise t-tests showed that this effect was significantly larger in the training group relative to

[1] It is worth noting that due to coding error, the cue-action mappings were not perfectly counterbalanced in the NoTraining+50% group (eight participants under one mapping and 16 under another). However, given that these subgroups behaved comparably, and that there were no effects in either of the 50% groups, we believe this error is unlikely to have changed the patterns in the data.

the group without a training phase ($t(46) = 2.87$, $p = .006$, $d = 0.83$). The magnitude of the RT expectation effect did not differ between the two 50% groups ($t(46) = 0.33$, $p = .740$, $d = 0.10$) and neither exhibited a significant expectation effect (Training+50%: $M = -3.46$ ms, $SD = 31.43$, $t(23) = -0.54$, $p = .595$, $d = -0.11$; NoTraining+50%: $M = -0.72$ ms, $SD = 24.99$, $t(23) = -0.14$, $p = .889$, $d = -0.03$). Replicating the findings from Experiment 1, the RT expectation effect was significantly larger in the Training+75% compared to the Training+50% group ($t(46) = 4.57$, $p < .001$, $d = 1.32$).

A two-way ANOVA on the PE expectation effect showed similar results. There was a significant main effect of both Degradation ($F(1,92) = 30.07$, $p < .001$, $\eta_p^2 = .25$) and Training ($F(1,92) = 12.24$, $p < .001$, $\eta_p^2 = .12$) as well as a significant interaction ($F(1,92) = 15.60$, $p < .001$, $\eta_p^2 = .15$; Fig. 2D). Both 75% groups showed a significant effect of expectations on PE (Training+75%: $M = 0.07$, $SD = 0.05$, $t(23) = 6.33$, $p < .001$, $d = 1.29$; NoTraining+75%: $M = 0.01$, $SD = 0.03$, $t(23) = 2.54$, $p = .018$, $d = 0.52$), and again this effect was larger in the training group ($t(46) = 4.41$, $p < .001$, $d = 1.27$). The magnitude of the PE expectation effect did not differ between the two 50% groups ($t(46) = 0.42$, $p = .675$, $d = 0.12$) and neither exhibited a significant expectation effect (Training+50%: $M = 0.0002$, $SD = 0.02$, $t(23) = 0.03$, $p = .975$, $d = 0.01$; NoTraining+50%: $M = 0.003$, $SD = 0.03$, $t(23) = 0.58$, $p = .567$, $d = 0.12$). Again replicating the findings from Experiment 1, the PE expectation effect was significantly larger in the Training+75% compared to the Training+50% group ($t(46) = 5.73$, $p < .001$, $d = 1.65$; see Supplementary Information for consideration of the development of these effects across trial).

These results reflect that expectation effects diminish as (previously strong) statistical relationships degrade – as observed in Experiment 1 and in the training groups of Experiment 2. Comparison against the no training groups in Experiment 2 renders unlikely an account whereby participants discarded expectations learned from training phases and relied solely on the statistics of the test phase.

4. Drift Diffusion Modelling

The present findings are consistent with Bayesian accounts of how expectations influence perceptual decisions, such that they reflect a precision-weighted combination of what we expect (prior) and the input (likelihood; Kersten et al., 2004) – with the prior reflecting the past and present statistics of the environment. Such influences can be generated in different ways and their nature can be revealed via drift diffusion modelling (DDM; Ratcliff & McKoon, 2008). DDMs model participant choices and RT distributions by assuming that observers have an internal representation of sensory evidence that is sampled by decision circuits. Decision circuits continuously sample from these representations, and when the accumulated decision variable meets a response boundary, the appropriate response is triggered (Fig. 3A and B). There are two ways that the observed expectation effects, and their degradation, could be generated. First, expectations could shift the starting point of the evidence accumulation process towards the expected boundary (varying parameter z of the DDM; Fig 3A). Second, expectations can bias the rate of evidence accumulation towards the two boundaries (varying parameter dc of the DDM; Fig. 3B). Our previous studies have found that drift rate biases explain influences of expectations on perceptual decisions (Thomas et al., 2022; Yon et al., 2021), in line with suggestions that expectations act by increasing the gain of related channels (de Lange et al., 2018; Summerfield & de Lange, 2014). We therefore anticipated that our effects of certainty degradation would be best modelled as modulations of the rate of evidence accumulation.

We fit hierarchical DDMs to participants' accuracy and RT data for Experiments 1 and 2 to examine how degradation effects are generated. We used the HDDM Python toolbox (Wiecki et al., 2013) to fit the DDMs. In the hierarchical DDM, model parameters for each participant are treated as random effects drawn from group-level distributions, and Bayesian Markov Chain Monte Carlo (MCMC) sampling is used to estimate group and participant level parameters simultaneously. We fit four different models to the data from each experiment: (1) a null model where no parameters were permitted to vary according to whether a CW or CCW

stimulus was expected; (2) a start bias model where the start point of evidence accumulation could vary between trials where a CW grating was expected (z_{CW}) and trials where a CCW grating was expected (z_{CCW}); (3) a drift bias model where a constant bias in evidence accumulation could vary between expected CW (dc_{CW}) and expected CCW (dc_{CCW}) trials; (4) a drift + start bias model where both parameters could vary. For Experiment 1, the parameters could additionally vary according to Degradation and for Experiment 2 according to Degradation and Training. This allowed us to obtain drift and start bias parameters for expected CW and expected CCW trials separately for each group in Experiments 1 and 2. Note that we use these terms to refer to the stimulus that is expected, while also modelling the stimulus that is presented – i.e., we include both expected and unexpected trial types but such coding allows us to examine influences of expectation. More specifically, larger start or drift bias values on expected CW compared to expected CCW trials would mean that participants are biased in line with their expectations (i.e., towards responding CW/CCW on an expected CW/expected CCW trial, respectively). Such biasing in line with participants' expectations would lead to faster and less error-prone responses on expected compared to unexpected trials (the behavioural effects we found in Experiments 1 and 2).

All models were estimated with MCMC sampling. For each model, three chains with 90,000 samples ("burn-in" = 22,500, "thinning" = 3) were run. This allowed us to compute the R-hat (Gelman-Rubin) statistic to assess model convergence. Typically, if this statistic is below 1.2 for all model parameters, one can be fairly confident that convergence has been reached (or even more reassuring is the more stringent condition < 1.1; Brooks & Gelman, 1998). For all experiments and all models, the Gelman-Rubin statistic was below 1.2 for all parameters, and below 1.1 for 99.92% of the parameters. Given that the Gelman-Rubin statistic indicated acceptable convergence, the chains of the three individual models were concatenated to create the final model.

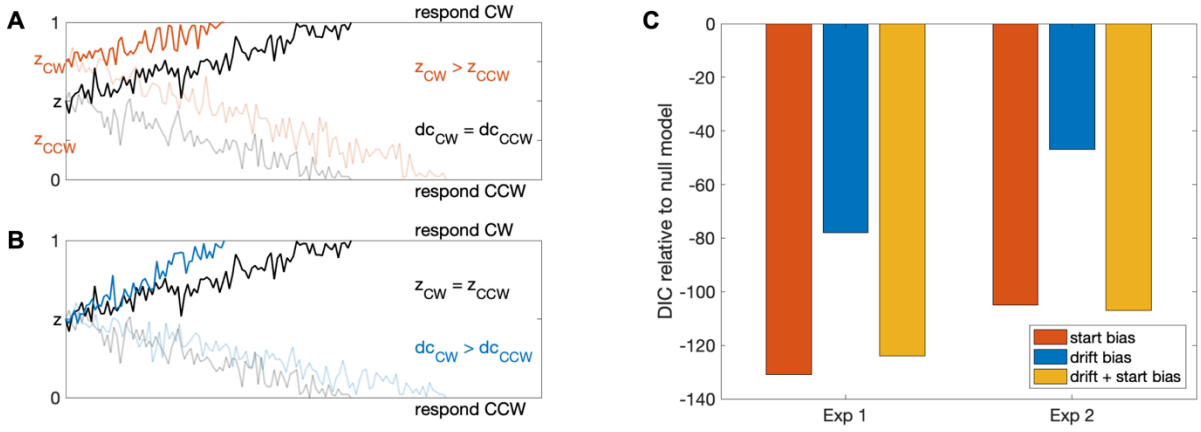

Figure 3. Drift diffusion modelling. (A) A schematic example of start biasing. The starting point of the evidence accumulation could be shifted towards the expected boundary from the outset of a trial, such that it starts closer to the CW response boundary when a CW grating is expected (z_{CW}), and closer to the CCW boundary when a CCW grating is expected (z_{CCW}). **(B)** A schematic example of drift biasing. Alternatively, the rate of evidence accumulation could be weighted according to expectations, such that when a CW grating is expected, evidence is accumulated more rapidly towards the CW boundary. **(C)** Comparison of the different models by deviance information criterion (DIC) values in Experiments 1 and 2. The drift bias model was outperformed by the start bias and drift + start bias models in both experiments. Lower DIC values relative to the null model are indicative of a better model fit.

The start, drift, and start + drift bias models for both experiments were compared using the deviance information criterion (DIC) as an approximation of Bayesian model evidence (Fig. 3C). Lower DIC values relative to the null model indicate better model fit. For both Experiments 1 and 2, the start bias and combined models outperformed the drift bias model (DIC relative to null model for Experiment 1: start bias model = -131, drift bias model = -78, start + drift bias model = -124; DIC relative to null model for Experiment 2: start bias model = -105, drift bias model = -47, start + drift bias model = -107). It is worth noting that differences in DIC smaller than 10 do not indicate substantial evidence for the model with the lower DIC (Dunovan et al., 2014; Zhang & Rowe, 2014). Given that the DIC difference between the start and start + drift bias models was less than 10 in both Experiment 1 and 2, we further analysed the parameters of the combined model in order to consider the impact of both start and drift biasing (noting also that our hypothesized patterns – based upon previous findings – pertained to drift biases and therefore we had reason to believe that start biases alone were unlikely to explain effects).

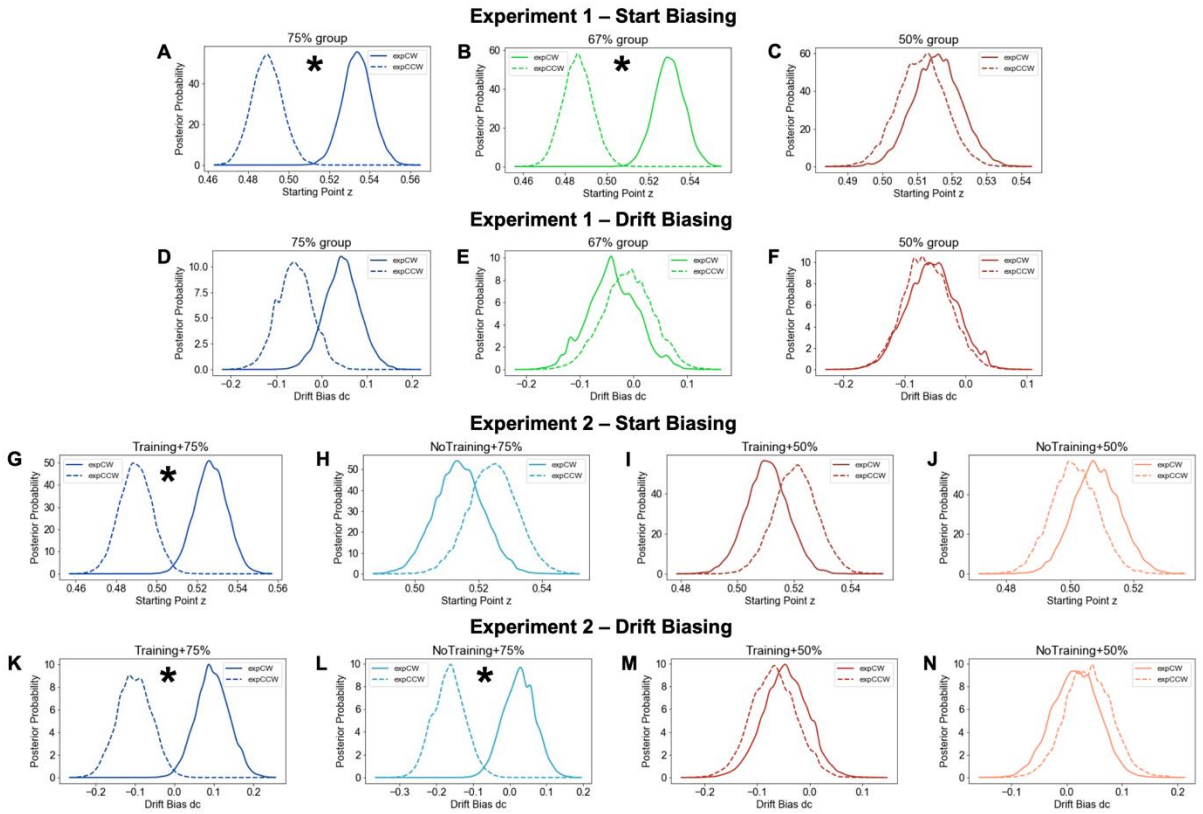

Figure 4. Posterior density plots of the modelled parameters from the start + drift bias model. The solid line shows the posterior density for the parameters on expected CW trials and the dashed line is for expected CCW trials. Stars mark significant differences. **(A-C)** Starting point for Experiment 1. The starting point for expected CW trials was larger than for expected CCW trials for the 75% and 67% groups, but not for the 50% group. **(D-F)** Drift bias for Experiment 1. There was no difference in drift bias for expected CW and expected CCW trials for any of the three groups, although the effect in the 75% group nearly met threshold. **(G-J)** Starting point for Experiment 2. The starting points differed on expected CW compared to expected CCW trials only for the Training+75% group. **(K-N)** Drift bias for Experiment 2. The drift bias was larger on expected CW compared to expected CCW trials for both 75% groups, but there were no significant differences for the 50% groups.

Specifically, we computed the proportion of overlap of the posterior distributions for expected CW and expected CCW trials for each group (see Fig. 4). Larger posterior probability values reflect smaller overlap in distributions (threshold for significance >.975). For Experiment 1, this revealed a starting point difference on expected CW compared to expected CCW trials for the 75% (posterior probabilities that $z_{CW} > z_{CCW} = 1.0$; Fig. 4A) and 67% ($z_{CW} > z_{CCW} = 1.0$; Fig. 4B) groups, but not for the 50% group ($z_{CW} > z_{CCW} = 0.679$; Fig. 4C). The drift bias for expected CW and expected CCW trials did not differ for any of the three groups (posterior probabilities that $dc_{CW} > dc_{CCW}$: 75% group = 0.967 [although note that this effect almost passed threshold],

67% group = 0.342, 50% group = 0.556; Fig. 4D-F). In Experiment 2, the start biases differed only for the Training+75% group (posterior probabilities that $z_{CW} > z_{CCW}$ = 0.999; Fig. 4G), but not for any of the other groups (posterior probabilities that $z_{CW} > z_{CCW}$: NoTraining+75% = 0.172, Training+50% = 0.174, NoTraining+50% = 0.747; Fig. 4H-J). In addition, the drift biases differed for both 75% groups (posterior probabilities that $dc_{CW} > dc_{CCW}$: Training+75% = 0.999, NoTraining+75% = 0.998; Fig. 4K-L), but for neither of the 50% groups (posterior probabilities that $dc_{CW} > dc_{CCW}$: Training+50% = 0.648, NoTraining+50% = 0.354; Fig. 4M-N).

6. General Discussion

In our ever-changing environment, it is common that we initially forge expectations based upon strong statistical relationships that subsequently degrade. The present studies were designed to interrogate how these altered statistics affect predictive influences on perceptual decisions. Across these studies we found evidence that we do not stubbornly, and suboptimally, maintain perceptual predictions established under strong statistical relationships. Instead, these results suggest that we update our expectations on the basis of new statistical relationships. Drift diffusion modelling indicated that these effects are explained by shifting the starting point in the evidence accumulation process as well as biasing the rate of evidence accumulation – with the former reflecting biases from statistics within the training session and the latter those of the test session.

Rapid expectation updating when the environment changes is likely adaptive for perception. Our perceptual experiences are constructed via interactions between the sensory input and our expectations (Bar, 2004; de Lange et al., 2018; Yuille & Kersten, 2006). It is reasoned that perceptual systems generate a precision-weighted sum of inputs and expectations to render experiences more accurate (on average) – given the internal and external noise associated with sensory signals and that we must process inputs rapidly in our ever-evolving sensory world (Press & Yon, 2019). If our expectations are based upon old knowledge, this process is less likely to generate veridical experiences. Of course, these mechanisms will only make

experiences more veridical *on average*, and expectations will frequently be unfulfilled. This is arguably why such strange experiences result when the typical contingencies are disrupted, as evidenced by a range of compelling illusions – e.g., the hollow face illusion, whereby a concave face is perceived as a (more typically-experienced) convex structure (Press, Kok, et al., 2020b).

Despite the fact that we observed a level of updating of perceptual expectations that might be considered adaptive, it was far from clear that we would obtain these results – which exhibit a greater optimality than might be concluded from previous work. For example, studies from our lab have observed that predictions persist when relationships disappear entirely (Press, Thomas, et al., 2020; Yon et al., 2018; Yon & Press, 2017, 2018) – extending even to scenarios where 'expected' events are presented on only 33% of trials (Press, Thomas, et al., 2020). A range of findings from animal learning also suggest completely unaltered predictions when statistical relationships degrade via extinction (e.g., Delamater, 2012; Rescorla, 1992, 1993). The present findings suggest that perceptual predictions based upon strong statistical relationships are not in fact supremely stubborn. Interestingly, our conclusions are supported also by more recent findings from animal learning that reward-learning associations *do* degrade inline with diminishing contingencies, and that the previous measures may not adequately reflect association strength (Crimmins et al., 2021). Recent accounts of goal-directed action control further suggest that we establish relationships according to a running rate-correlation between actions and outcomes (Perez & Dickinson, 2020), and therefore that extinction may represent a rather unique case of statistical degradation.

The fact that no biasing is observed at all when previously perfect relationships disappear suggests that expectations *can be* especially flexible. However, these findings must also be considered in light of our previous work demonstrating that predictions can persist when correlations vanish. Most of these findings target priors established over a different timescale – specifically, a lifetime's experience of perceiving particular outcomes of action (Yon et al., 2021; Yon & Press, 2017, 2018). It is possible that expectations established over different

timescales exhibit differing levels of ‘stubbornness’. Those established over one’s lifetime are less likely to be found incorrect based upon a few hundred trials, relative to those established within a one-hour period. In other words, a comparable correlation based upon more datapoints will generate an even higher precision prior, and will therefore require even stronger evidence to update. This difference is consistent with the suggestion that you may be able to extinguish expectations that are ‘weakly’ learnt but not ‘strong’ ones (Delamater, 2012).

Comparing the 75% groups in Experiment 2 demonstrates that the biasing influences on perception are substantially stronger when there is a preceding 100% training phase, relative to when expectations could only be established within the test session. Interestingly these training vs no training group differences are not present in the 50% groups, suggesting that eliminating predictive relationships can cause the individual to discard earlier learning in a way that reducing them does not. However, future research must also establish whether learning is discarded by all systems when correlations disappear, or whether there may be differences across them. A recent fMRI study of ours (Press, Thomas, et al., 2020) suggested that primary visual cortex maintains a representation of relationships learned the previous day even when the ‘expected’ is in fact unlikely (presented only 33% of the time). Participants here were subjected to more training than in the present studies, but it may also suggest that primary visual cortex can maintain a representation based upon old knowledge that has little influence on perceptual decisions. Future work should focus on the interplay between neural and decisional effects to determine how they may inform each other. There are also some clear practical implications of this finding. It is common to examine the influence of expectations on perceptual processing by presenting the relationships solely at test – in degraded form. Our findings indicate that effect sizes would be larger if including preceding training phases, enabling greater clarity in findings and/or smaller sample sizes.

If we indeed discard previous learning when correlations disappear it would be interesting to understand why, and the empirical limits. Speculatively, individuals may ascribe different latent structures to training and test phases when correlations disappear, but not when they remain

in degraded form (Niv, 2019). Participants may typically assume that the environment remains stable, but there may have been meta-cues in the 50% group indicating that the structure had altered across sessions. Although participants were given identical instructions on the two days that did not reference action-outcome mappings (and identical instructions in 50% and 75% groups), alongside a short perfectly predictive training refresher at the start of the second session in Experiment 2, we do not know what assumptions they may have built independently. It would be informative to ask about the limits at which one discards previous learning – e.g., whether a 67% training vs no training comparison would reveal less discard than the 50% and more than the 75% comparison, and whether learning is discarded in the same manner when contingencies degrade gradually. In line with the operation of some all-or-none mechanisms, some of us have previously proposed that qualitatively different processes may influence perception when the level of prediction error crosses a threshold – which elicits reactive processes to generate a high-precision representation of events which should generate model-updating (Press, Kok, et al., 2020b).

Drift diffusion modelling indicated that effects of correlation degradation were the product of shifting the starting point of the accumulation process as well as biasing the rate of evidence accumulation towards the expected boundary. Start point effects were unanticipated, given our previous studies that found expectations based upon action congruency – therefore reflective of a lifetime of learning about the consequences of action – modulated the rate of evidence accumulation (Yon et al., 2021). However, they are less surprising if one considers recent suggestions about the particular mechanisms underlying influences of expectation on perceptual decisions (Feuerriegel et al., 2021). Specifically, it has been suggested that they act by pre-activating sensory representations of anticipated events (Gandolfo & Downing, 2019; Kok et al., 2017; Press, Kok, et al., 2020b; de Lange et al., 2018). This mechanism shifts the evidence accumulation process earlier – to before the presentation of the stimulus – and it has been proposed that such mechanisms manifest in different ways depending on specifics of the stimuli presented (Feuerriegel et al., 2021). For instance, it has been argued that when

the input itself is weak or punctate (like in Yon et al., 2021), drift rate effects should emerge from shifting the accumulation process earlier because the rate of evidence accumulation grows across time. However, Feuerriegel et al. (2021) argue that when the stimulus is presented for longer (like in the present studies), the drift rate is largely determined by the stimulus and pre-activation presents predominantly as a shift in the start point of evidence accumulation.

Experiment 2 found that the particular degradation level at test instead influenced the drift rate rather than starting point – in spite of the longer duration events (noting the above proposal by Feuerriegel et al., 2021, that event durations largely determine the observed effects from pre-activation mechanisms). Specifically, both training and no training 75% groups in Experiment 2 showed drift biasing effects whereas only the training 75% group showed start biasing effects of expectation. An interesting factor that could determine this difference is that of consolidation. For instance, the start biasing effects could reflect learning during the training session on the previous day (which had been consolidated), and therefore biases us to perceive what we expect via pre-activation. On the other hand, drift biasing differences may reflect learning across the test session – and it is possible that the underlying mechanisms differ accordingly. For example, pre-activation may require learning consolidation via sleep (Hindy et al., 2019) or the presentation of high predictive relationships.

Future research is required to unpack the relative contributions of sensory and purely decisional mechanisms to these phenomena (Fritsche et al., 2017). Particularly, start point biases in decision processes are classically assumed (including by us) to be reflective of decisional biasing rather than sensory biasing, because they are independent of the evidence accumulated (Yon et al., 2021). This remains a viable possibility on the basis of the present dataset. However, if expectations bias perception by pre-activating associated sensory representations, this will also shift the starting point of the evidence accumulation process (Feuerriegel et al., 2021). If this is indeed how the system operates, the present findings suggest that some level of nuance is needed in these mechanisms – either scaling pre-

activations according to likelihood of event presentation or pre-activating multiple representations simultaneously.

In conclusion, the present data demonstrate how degrading statistical relationships degrades perceptual expectations. Even if perceptual expectations are often more stubborn than they should be, the present study shows how we do update them to inform our perceptual decisions when statistical relationships become less certain.

Acknowledgements

We are grateful to Milan Andrejevic, Anthony Dickinson, Mark Haselgrove and Daniel Yon for useful discussions, to Matan Mazor and Emma Ward for comments on the manuscript, and to Jessica Nicholson for assistance with Experiment 1. The work was supported by a Leverhulme Trust project grant (RPG-2016-105) and European Research Council consolidator grant (101001592) under the European Union's Horizon 2020 research and innovation programme, both awarded to CP.

Author contributions

Conceptualization: ET, KR and CP. Formal analysis, investigation, and project administration: ET and KR. Writing – original draft: ET. Writing – review and editing: KR and CP.

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Raincloud plots: A multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. *Wellcome Open Research*, *4*, 63. https://doi.org/10.12688/wellcomeopenres.15191.1

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Crimmins, B., McNulty, M., Laurent, V., Hart, G., & Balleine, B. (2021). *Response-independent outcome presentations weakens the instrumental response-outcome association*. PsyArXiv. https://doi.org/10.31234/osf.io/c83kx

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

Delamater, A. R. (2012). Issues in the extinction of specific stimulus-outcome associations in Pavlovian conditioning. *Behavioural Processes*, *90*(1), 9–19. https://doi.org/10.1016/j.beproc.2012.03.006

Dickinson, A., & Charnock, D. J. (1985). Contingency effects with maintained instrumental reinforcement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(4b), 397–416. https://doi.org/10.1080/14640748508401177

Dunovan, K. E., Tremel, J. J., & Wheeler, M. E. (2014). Prior probability and feature predictability interactively bias perceptual decisions. *Neuropsychologia*, *61*, 210–221. https://doi.org/10.1016/j.neuropsychologia.2014.06.024

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Feuerriegel, D., Blom, T., & Hogendoorn, H. (2021). Predictive activation of sensory representations as a source of evidence in perceptual decision-making. *Cortex*, *136*, 140–146. https://doi.org/10.1016/j.cortex.2020.12.008

Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite effects of recent history on perception and decision. *Current Biology*, *27*(4), 590–595. https://doi.org/10.1016/j.cub.2017.01.006

Gandolfo, M., & Downing, P. E. (2019). Causal evidence for expression of perceptual expectations in category-selective extrastriate regions. *Current Biology*, *29*(15), 2496-2500.e3. https://doi.org/10.1016/j.cub.2019.06.024

Garrido, M. I., Sahani, M., & Dolan, R. J. (2013). Outlier responses reflect sensitivity to statistical structure in the human brain. *PLOS Computational Biology*, *9*(3), e1002999. https://doi.org/10.1371/journal.pcbi.1002999

Hindy, N. C., Avery, E. W., & Turk-Browne, N. B. (2019). Hippocampal-neocortical interactions sharpen over time for predictive actions. *Nature Communications*, *10*(1), 3989. https://doi.org/10.1038/s41467-019-12016-9

Johansson, R. S., Westling, G., Bäckström, A., & Flanagan, J. R. (2001). Eye–hand coordination in object manipulation. *Journal of Neuroscience*, *21*(17), 6917–6932. https://doi.org/10.1523/JNEUROSCI.21-17-06917.2001

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.*, *55*, 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Kok, P., Mostert, P., & Lange, F. P. de. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences*, *114*(39), 10473–10478. https://doi.org/10.1073/pnas.1705652114

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8

Oaksford, M., & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. OUP Oxford.

Ouden, H. E. M. den, Daunizeau, J., Roiser, J., Friston, K. J., & Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *Journal of Neuroscience*, *30*(9), 3210–3219. https://doi.org/10.1523/JNEUROSCI.4458-09.2010

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526. https://doi.org/10.3758/BF03197524

Perez, O. D., & Dickinson, A. (2020). A theory of actions and habits: The interaction of rate correlation and contiguity systems in free-operant behavior. *Psychological Review*, *127*(6), 945–971. https://doi.org/10.1037/rev0000201

Press, C., Kok, P., & Yon, D. (2020a). Learning to perceive and perceiving to learn. *Trends in Cognitive Sciences*, *24*(4), 260–261. https://doi.org/10.1016/j.tics.2020.01.002

Press, C., Kok, P., & Yon, D. (2020b). The perceptual prediction paradox. *Trends in Cognitive Sciences*, *24*(1), 13–24. https://doi.org/10.1016/j.tics.2019.11.003

Press, C., Thomas, E., Gilbert, S., Lange, F. de, Kok, P., & Yon, D. (2020). Neurocomputational mechanisms of action-outcome prediction in V1. *Journal of Vision*, *20*(11), 712–712. https://doi.org/10.1167/jov.20.11.712

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., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Rescorla, R. A. (1992). Response-independent outcome presentation can leave instrumental R-O associations intact. *Animal Learning & Behavior*, *20*(2), 104–111. https://doi.org/10.3758/BF03200407

Rescorla, R. A. (1993). Preservation of response-outcome associations through extinction. *Animal Learning & Behavior*, *21*(3), 238–245. https://doi.org/10.3758/BF03197988

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning: Current research and theory*.

Skora, L. I., Seth, A. K., & Scott, R. B. (2021). Sensorimotor predictions shape reported conscious visual experience in a breaking continuous flash suppression task. *Neuroscience of Consciousness*, *2021*(1). https://doi.org/10.1093/nc/niab003

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756. https://doi.org/10.1038/nrn3838

Thomas, E. R., Yon, D., de Lange, F. P., & Press, C. (2022). Action enhances predicted touch. *Psychological Science*, *33*(1), 48–59. https://doi.org/10.1177/09567976211017505

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*. https://doi.org/10.3389/fninf.2013.00014

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., & Press, C. (2014). Back to the future: Synaesthesia could be due to associative learning. *Frontiers in Psychology*, *5*, 702. https://doi.org/10.3389/fpsyg.2014.00702

Yon, D., & Press, C. (2017). Predicted action consequences are perceptually facilitated before cancellation. *Journal of Experimental Psychology. Human Perception and Performance*, *43*(6), 1073–1083. https://doi.org/10.1037/xhp0000385

Yon, D., & Press, C. (2018). Sensory predictions during action support perception of imitative reactions across suprasecond delays. *Cognition*, *173*, 21–27. https://doi.org/10.1016/j.cognition.2017.12.008

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

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

Zhang, J., & Rowe, J. (2014). Dissociable mechanisms of speed-accuracy tradeoff during visual perceptual learning are revealed by a hierarchical drift diffusion model. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00069