Cognition 192 (2019) 104004

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Predictive cues reduce but do not eliminate intrinsic response bias

Mingjia Hu, Dobromir Rahnev*

School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Perceptual decision making
Bias
Criterion
Prediction
Signal detection theory

ABSTRACT

Predictive cues induce large changes in people's choices by biasing responses towards the expected stimulus category. At the same time, even in the absence of predictive cues, humans often exhibit substantial intrinsic response biases. Despite the ubiquity of both of these biasing effects, it remains unclear how predictive cues interact with intrinsic bias. To understand the nature of this interaction, we examined data across three previous experiments that featured a combination of neutral cues (revealing intrinsic biases) and predictive cues. We found that predictive cues decreased the intrinsic bias to about half of its original size. This result held both when bias was quantified as the criterion location estimated using signal detection theory and as the probability of choosing a particular stimulus category. Our findings demonstrate that predictive cues reduce but do not eliminate intrinsic response bias, testifying to both the malleability and rigidity of intrinsic biases.

1. Introduction

Perceptual decision making is the process of making a judgment about the identity of a stimulus based on the available sensory information (Hanks & Summerfield, 2017). Perceptual decisions can be described as a combination of two quantities: stimulus sensitivity, which quantifies a subject's intrinsic ability to perform the task, and response bias, which quantifies a subject's propensity to choose one stimulus category over another (Green & Swets, 1966). Stimulus sensitivity and response bias together determine how a subject responds to any given stimulus. However, although stimulus sensitivity has been the object of intense study, response bias remains poorly understood.

Two main sources of response bias have been described in the literature. First, response bias can be manipulated experimentally by providing unequal priors or rewards (Fig. 1A) (Ackermann & Landy, 2015; Bohil & Maddox, 2001, 2003; Busemeyer & Myung, 1992; Healy & Kubovy, 1981; Maddox & Bohil, 1998; Maddox & Dodd, 2001; Maddox, 2002). Here we specifically focus on manipulations in which predictive cues indicate on a trial-by-trial basis the stimulus category that is more likely to occur (Cheadle, Egner, Wyart, Wu, & Summerfield, 2015; Jiang, Summerfield, & Egner, 2013; Kok, Mostert, & de Lange, 2017; Kok, Rahnev, Jehee, Lau, & De Lange, 2012; Morales et al., 2015; Todorovic & de Lange, 2012). Such predictive cues change subjects' priors for the two stimulus categories and induce a corresponding bias in subjects' responses (for reviews, see Rahnev & Denison, 2018; Summerfield & de Lange, 2014).

The other source of response bias is the intrinsic propensity of subjects to prefer one stimulus category over another (Fig. 1B). Humans are known to have stable idiosyncratic biases that differ from subject to subject (Finlayson, Papageorgiou, & Schwarzkopf, 2017; García-Pérez & Alcalá-Quintana, 2011; Kosovicheva & Whitney, 2017; Linares, Aguilar-Lleyda, & López-Moliner, 2019; Wexler, Duyck, & Mamassian, 2015). We recently showed that these biases are robust across multiples days of the same experiment and extend to even the simplest two-choice perceptual tasks (Rahnev & Denison, 2018).

Although the existence of biases induced by predictive cues or intrinsic factors is well established, it has remained unclear how these two sources of bias interact with each other. Specifically, three alternative possibilities could potentially describe this interaction.

First, predictive cues may completely eliminate the intrinsic bias. We call this the "eliminate" hypothesis; it corresponds to the intuition that intrinsic response biases are malleable and could be eliminated when external priors are presented. This hypothesis predicts that the average of the criteria corresponding to the cues predicting each stimulus category would center on no-bias criterion location (Fig. 1C).

Second, predictive cues may maintain the intrinsic bias and simply stretch it to accommodate the new information regarding priors. We call this the "maintain" hypothesis; it corresponds to the intuition that the intrinsic response bias may be rigid and thus could become the new baseline around which any experimental manipulations operate. This hypothesis predicts that the average of the criteria corresponding to the cues predicting each stimulus category would center on criterion location for the intrinsic bias (Fig. 1D).

Third, predictive cues may reduce but not eliminate the intrinsic

* Corresponding author at: Georgia Institute of Technology, 654 Cherry Str NW, Atlanta, GA 30332, USA.
E-mail address: rahnev@psych.gatech.edu (D. Rahnev).

https://doi.org/10.1016/j.cognition.2019.06.016
Received 18 December 2018; Received in revised form 13 June 2019; Accepted 14 June 2019

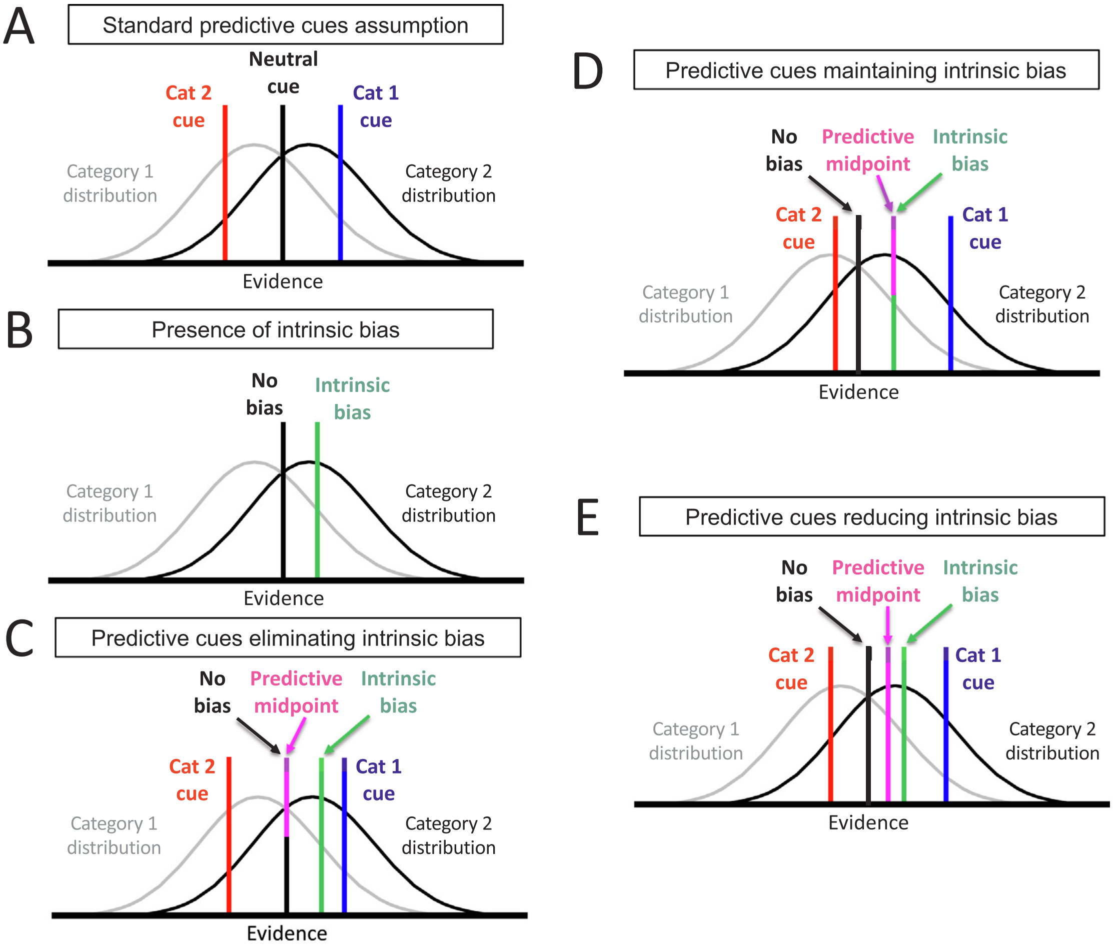

Fig. 1. Interaction between predictive cues and intrinsic response bias. Gaussian stimulus distribution for stimuli from Categories 1 and 2 are displayed in gray and black, respectively. (A) Standard predictive cues assumption. A neutral cue with no predictive value is typically assumed to result in an unbiased criterion placed at the intersection of the two Gaussian distributions. Predictive cues indicating that Category 1 or Category 2 is more likely to occur are assumed to lead to a symmetric shift in the criterion location around the neutral cue criterion. (B) Presence of intrinsic bias. Even though subjects *should* place their criterion at the intersection of the two Gaussian distributions, it is often the case that intrinsic biases shift the criterion away from that location. In the case depicted here, the intrinsic bias favors Category 1. (C) "Eliminate" hypothesis: predictive cues eliminate intrinsic response bias. If so, the midpoint of the two predictive criteria (magenta) overlaps with the no-bias criterion (black). (D) "Maintain" hypothesis: predictive cues maintain intrinsic response bias. If so, the midpoint of the two predictive criteria (magenta) overlaps with the intrinsic response bias criterion (green). (E) "Reduce" hypothesis: predictive cues reduce intrinsic response bias. If so, the midpoint of the two predictive criteria falls in-between the no-bias and intrinsic bias criteria. Cat 1, Category 1; Cat 2, Category 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

bias. We call this the "reduce" hypothesis; it corresponds to the intuition that the intrinsic response bias may be both malleable and rigid. This hypothesis predicts that the average of the criteria corresponding to the cues predicting each stimulus category would typically fall in-between the no-bias and intrinsic bias criterion locations (Fig. 1E).

In order to adjudicate between these competing hypotheses, we re-analyzed data from three previous experiments (Bang & Rahnev, 2017; de Lange, Rahnev, Donner, & Lau, 2013; Rahnev, Lau, & De Lange, 2011). Each experiment featured an expectation manipulation such that, on a trial-by-trial basis, a predictive cue indicated which stimulus category is more likely to occur. Three types of cues were used in each case: cues predicting the first stimulus category, cues predicting the second stimulus category, and neutral cues with no predictive value. This design allowed us to investigate how biases induced by each of the two predictive cues interact with subjects' intrinsic biases in the neutral cue condition. To anticipate, the results strongly supported the "reduce" hypothesis: Predictive cues reduced (by about half) but did not eliminate intrinsic response bias.

2. Methods

2.1. Subjects

A total of 72 healthy subjects participated in the three experiments (30 subjects in Experiment 1, 23 subjects in Experiment 2, and 19 subjects in Experiment 3). All three experiments were previously published (Bang & Rahnev, 2017; de Lange et al., 2013; Rahnev et al., 2011). In Experiment 2, two subjects were excluded in the original publication due to chance-level performance; the same subjects were excluded in the analyses here too. The original sample sizes were outside of our current control and were individually too small for reliable parameter estimates in the context of our Binomial and regression tests. Therefore, we combined the data from the three experiments, giving us a total of 70 subjects. This sample allows us to detect even relatively small effect sizes. For completeness, we still report the results of all analyses as run on each individual experiment. All subjects had normal or corrected-to-normal vision and gave written informed consent. All experiments were approved by the local Institutional Review Board.

2.2. Experiment selection

We analyzed all relevant experiments for which we had access to the raw data. Because each of these experiments had a relatively small sample size ($N \leq 30$), we conducted a comprehensive search for published papers that contain data relevant to our current analyses. However, this search failed to produce any studies with sample sizes larger than the experiments already included here (sample size range = 6–23). Therefore, we elected to only analyze the data from Bang and Rahnev (2017), Rahnev et al. (2011), and de Lange et al. (2013) for which we already had the raw data.

2.3. Commonality in experimental design for experiments 1–3

The three experiments differed in a number of dimensions. However, they were chosen for re-analysis here because of the commonality between them. Specifically, each experiment featured a combination of three cues: a cue predicting the first stimulus category, a cue predicting the second stimulus category, and a neutral cue with no predictive value. The predictive cues were valid in 66.67% of the trials in Experiment 1, and in 75% of trials in Experiments 2 and 3. The neutral cue always signified that each stimulus category is equally likely to occur. The presence of these three cue types allowed us to examine the interaction of the bias induced by the predictive cues with the bias observed for neutral cues. Full experimental details can be found in the original publications; below we highlight the most relevant design characteristics of each experiment.

2.4. Experiment 1 design

Experiment 1 was originally reported in Bang and Rahnev (2017). Subjects' task was to indicate whether the overall direction of a series of Gabor patches had an overall tilt to the left (i.e., clockwise) or to the right (i.e., counterclockwise) from the vertical. Each trial featured a sequential presentation of a pre cue, a fixation cross, a stimulus, a fixation cross, a post-cue (each phase lasted 500 ms), and an untimed response period (Fig. 2A). The stimulus consisted of 30 Gabor patches with orientations sampled randomly from a normal distribution with a mean determined separately for each subject (average = +/−7.49°) and a standard deviation of 22.5°. Each individual Gabor patch was presented for one computer frame (16.7 ms). After each trial, subjects received feedback on the accuracy of their response.

The predictive cues consisted of the following symbols: " < " indicated that an overall left tilt was more likely, " > " indicated that an overall right tilt was more likely, and "|" indicated that both tilts were equally likely. The original purpose of this experiment was to compare the influence of cues presented before or after the stimulus presentation. Therefore, the experiment featured "pre cue" and "post cue" blocks where the cues were presented before and after the stimulus, respectively. To keep the timing between these blocks consistent, an uninformative dash sign was presented after the stimulus in pre cue blocks and before the stimulus in post cue blocks.

Subjects completed a total of 480 trials each. Left/right/neutral cues were presented with 37.5%/37.5%/25% probability resulting in 180/180/120 trials total for each cue type. These trials were equally split between pre and post cue blocks. For the purposes of the present analyses, the pre and post cues were combined together. However, control analyses not shown here demonstrated that all results remain the same if only pre or post cues are considered.

2.5. Experiment 2

This study was originally reported in Rahnev et al. (2011). Subjects indicated the direction of motion (either contracting or expanding) of white dots (density, 2.4 dots/degree^2; speed, 6°/s) presented on a black annulus (outer circle radius, 10°; inner circle radius, 1°). The black annulus was clearly visible since it was presented on a gray background. Each trial featured a sequential presentation of a cue (250 ms), a fixation square (250 ms), a stimulus (500 ms), and a response prompt (1 s) (Fig. 2B). Subjects completed the task as part of a functional MRI experiment. Threshold motion coherence was determined based on a training session (mean = 4.4%, SD = 0.7%). The actual experiment featured equal number of three motion coherence levels: 50%, 100%, and 150% of the threshold coherence level. The three motion coherence levels were combined in the present analyses.

The cues were four geometric shapes (square, diamond, triangle

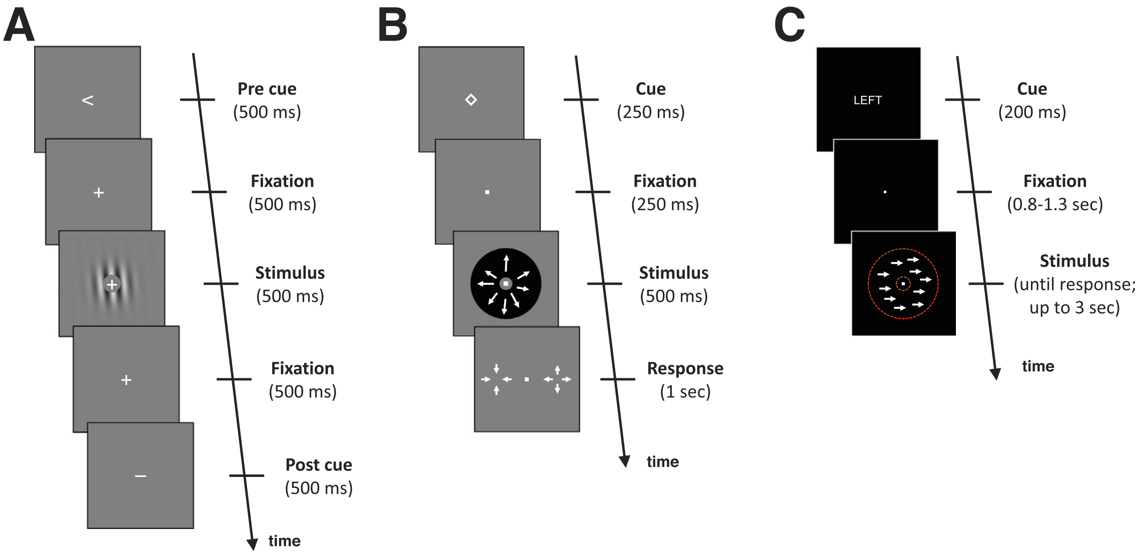

Fig. 2. Experimental tasks. All experiments featured a combination of cues predicting the first stimulus category, the second stimulus category, and neutral cues. Subjects' task was to indicate the stimulus category (1 or 2) that the stimulus came from. (A) Experiment 1 task. Predictive (66.67% validity) and neutral (no predictive validity) cues indicated the likely stimulus orientation (clockwise/counterclockwise). In different blocks, the predictive cue was presented before (pre cue) or after (post cue) the stimulus. These conditions were combined in the present analyses. (B) Experiment 2 task. Predictive (75% validity) and neutral (no predictive validity) cues indicated the likely motion direction (contracting/expanding). On each trial, the response mapping was indicated after the stimulus offset. (C) Experiment 3 task. Predictive (75% validity) and neutral (no predictive validity) cues indicated the likely motion direction (left/right). In all experiments, the stimulus difficulty was individually adjusted for each subject.

pointing up, and triangle pointing down). For half of the subjects the square and diamond were predictive (whereas both the triangles were non-predictive) and for the other half this relationship was reversed. Unlike in the other experiments, response mapping was provided only after the offset of the stimulus. Subjects responded with the index fingers of their left and right hands and the response mapping was indicated by a set of arrows pointing inward indicating the finger mapped to contracting motion, and a set of arrows pointing outward indicating the finger mapped to expanding motion.

Subjects completed a total of 672 trials. The experiment was organized in alternating 8-trial blocks of predictive and non-predictive cues. Predictive cue blocks featured cues predicting both stimulus categories. Contracting/expanding/neutral cues were presented with 25%/25%/50% probability resulting in 168/168/336 trials total for each cue type.

2.6. Experiment 3

This study was originally reported in de Lange et al. (2013). Subjects indicated the direction of motion (either left or right) of white dots (density, 2.4 dots/degree^2; speed, 6°/s) presented on a black annulus (outer circle radius, 10°; inner circle radius, 1°). The black annulus itself was not visible because the background was also black. Each trial featured a sequential presentation of a cue (200 ms), a fixation dot (800–1,300 ms), and a stimulus (presented until response, up to 3 s) (Fig. 2C). Subjects completed the task as part of a magnetoencephalography experiment. As in Experiment 2, three motion coherence levels were used and were combined for data analyses. The cues were the words "LEFT," "RIGHT," and "NEUTRAL."

Subjects completed a total of 864 trials. Left/right/neutral cues were presented with 33.3%/33.3%/33.3% probability resulting in 288/288/288 trials total for each cue type. The three types of cues were randomly interleaved.

2.7. Bias measures

To quantify subjects' tendency to respond in favor of one stimulus category, we calculated the signal detection theory (SDT) measure decision criterion (c) by calculating hit rate (HR) and false alarm rate (FAR):

$$c = -\frac{1}{2}\times[\Phi^{-1}(HR) + \Phi^{-1}(FAR)]$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution that transforms HR and FAR into z-scores. HR and FAR were defined by treating right orientations in Experiment 1, expanding motion in the Experiment 2, and rightward motion in the Experiment 3 as targets. In other words, Category 1/Category 2 corresponded to left/right orientation in Experiment 1, contracting/expanding motion in Experiment 2, and leftward/rightward motion in Experiment 3. Note that negative criterion c values indicate a bias for Category 2 stimuli, whereas positive criterion c values indicate a bias for Category 1 stimuli.

For each subject in each experiment, we computed the criteria used in the presence of Category 1 cues (c_{cat1}), Category 2 cues (c_{cat2}), and neutral cues ($c_{neutral}$). We then determined the midpoint of the criteria associate with the two predictive cues: $c_{PredMid} = \frac{c_{cat1} + c_{cat2}}{2}$ (see Supplementary Methods for why the midpoint of the two criteria represents the overall bias in the presence of predictive cues).

To check for the robustness of our results, we repeated these analyses with an alternative way of estimating response bias. In these control analyses, we defined response bias as the percent of "Category 1" responses, $P(\text{resp} = \text{Cat1})$. For this alternative measure of bias, we again computed the bias in the presence of Category 1 cues ($P_{cat1}(\text{resp} = \text{Cat1})$), Category 2 cues ($P_{cat2}(\text{resp} = \text{Cat1})$), and neutral cues ($P_{neutral}(\text{resp} = \text{Cat1})$). We then determined the midpoint of the biases associated with the two predictive cues: $P_{PredMid}(\text{resp} = \text{Cat1}) = \frac{P_{cat1}(\text{resp} = \text{Cat1}) + P_{cat2}(\text{resp} = \text{Cat1})}{2}$.

2.8. Binomial tests

We performed Binomial tests to determine whether the midpoint of the predictive criteria, $c_{PredMid}$, was centered around c_{NoBias} (as predicted by the "eliminate" hypothesis) or $c_{neutral}$ (as predicted by the "maintain" hypothesis). To do so, we determined the number of subjects for who the midpoint of the predictive criteria fell in one of three locations. First, $c_{PredMid}$ could fall on the opposite side of c_{NoBias} compared to $c_{neutral}$, that is $\begin{cases} c_{PredMid} < c_{NoBias},\ c_{neutral} \geq 0 \\ c_{PredMid} > c_{NoBias},\ c_{neutral} < 0 \end{cases}$. Second, $c_{PredMid}$ could fall in-between c_{NoBias} and $c_{neutral}$. Third, $c_{PredMid}$ could fall beyond $c_{neutral}$, that is $\begin{cases} c_{PredMid} > c_{neutral},\ c_{neutral} \geq 0 \\ c_{PredMid} < c_{neutral},\ c_{neutral} < 0 \end{cases}$. We then tested whether $c_{PredMid}$ was centered around c_{NoBias} by performing a two-sided Binomial test to examine whether the number of subjects in location 1 is equal to the sum of the number of subjects in locations 2 and 3. Similarly, we tested whether $c_{PredMid}$ was centered around $c_{neutral}$ by performing a two-sided Binomial test to examine whether the number of subjects in locations 1 and 2 is equal to the sum of the number of subjects in location 3. Equivalent analyses were also performed with the alternative measure of bias $P(\text{resp} = \text{Cat1})$.

The Binomial tests are less powerful than the regression analyses (discussed below) but we report them for two main reasons: (1) they can be used to quantify the number of subjects consistent with each hypothesis, and (2) they are less sensitive to the results of subjects with extreme intrinsic biases, thus decreasing any potential biasing influence by these subjects. We note that small sensitivity to the subjects with large biases is not necessarily an asset because it is hard to judge how predictive cues modulate the intrinsic bias if the bias is very small to begin with. Nevertheless, we see the binomial tests as a useful complementary approach that weighs each subject equally.

2.9. Regression analyses

We further tested the three hypotheses about the interaction between predictive cues and intrinsic response bias by performing a regression where $c_{PredMid}$ was predicted by $c_{neutral}$. The different hypotheses made different predictions about the β value obtained in this regression. The "eliminate" hypotheses, which states that predictive cues eliminate the intrinsic bias, implies that $c_{PredMid}$ should not be affected by the value of $c_{PredMid}$ (because the intrinsic bias has been eliminated) and therefore predicts that $\beta = 0$. The "maintain" hypotheses, which states that predictive cues maintain the intrinsic bias, implies that $c_{PredMid} = c_{neutral}$ (with added estimation noise) and therefore predicts that $\beta = 1$. Finally, the "reduce" hypotheses, which states that predictive cues reduce but do not eliminate intrinsic bias, implies that $c_{PredMid}$ falls in-between 0 and $c_{neutral}$ (again, with added estimation noise) and therefore predicts that $0 < \beta < 1$. To adjudicate between the different hypotheses, we compared the obtained β values with 0 and 1 using F tests.

2.10. The criterion depictions in Fig. 3

Fig. 3 plots the Gaussian distributions and criterion locations for two subjects using standard signal detection theory (SDT) assumptions. Standard SDT postulates that the two stimulus categories give rise to Gaussian distributions $N(\mu_{cat1}, \sigma^2)$ and $N(\mu_{cat2}, \sigma^2)$, respectively. The SDT parameter d' is then equal to $d' = \frac{\mu_{cat2} - \mu_{cat1}}{\sigma}$. All computations therefore remain the same if both μ's and σ are scaled by the same number, or if a constant is added to the μ's for each stimulus category. Therefore, without loss of generality, we set $\sigma = 1$ and $\mu_{cat1} = -\mu_{cat2}$. This particular decision has the property that it makes the x axis of evidence coincide precisely with the value of the criterion c. The plots in Fig. 3 were thus produced by, for each subject, drawing Gaussian

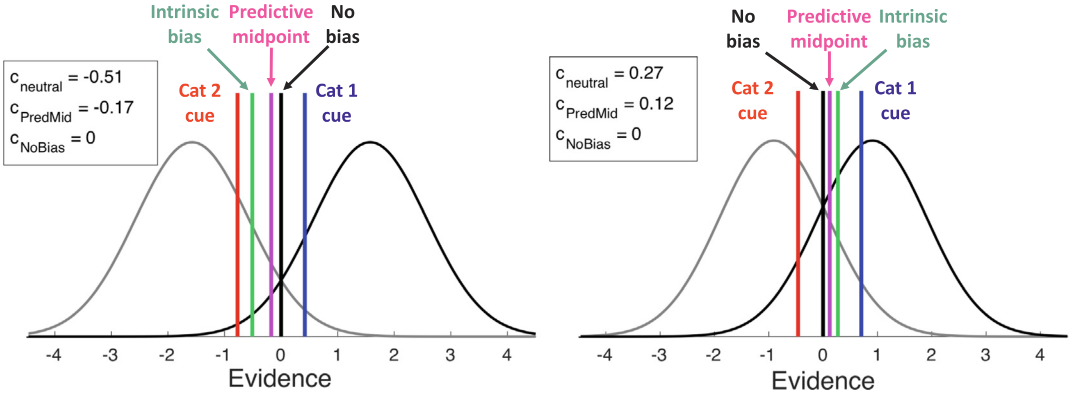

Fig. 3. Bias effects for two subjects. For illustrative purposes, we depict one subject with an intrinsic bias for responding "Category 2" (that is, with a negative neutral criterion; left panel) and one subject with an intrinsic bias for responding "Category 1" (that is, with a positive neutral criterion; right panel). For both subjects, the midpoint of the predictive criteria ($c_{PredMid}$) is located in-between the no-bias criterion (c_{NoBias}) and the intrinsic bias criterion ($c_{neutral}$). The insets list the exact values for each of these three criteria for each subject. The separation of the Gaussian distributions for each subject reflects that particular subject's stimulus sensitivity *d'*. Cat 1, Category 1; Cat 2, Category 2.

distributions with means $\pm \frac{d'}{2}$ and standard deviations of 1. Finally, *d'* was computed directly from subjects' data using the formula $d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR)$.

2.11. Supplementary data

All data and codes for the analyses have been made freely available at https://osf.io/xe8b3/ (DOI: https://doi.org//10.17605/OSF.IO/XE8B3).

3. Results

We investigated how human subjects' intrinsic response bias in perceptual decision making interacts with the response biases induced by predictive cues. Specifically, we adjudicated between three competing hypotheses: that predictive cues (1) eliminate, (2) maintain, or (3) reduce intrinsic response bias. We quantified each subject's bias as the signal detection theory (SDT) criterion in the presence of predictive and neutral cues across three different previously published experiments (Bang & Rahnev, 2017; de Lange et al., 2013; Rahnev et al., 2011).

Fig. 3 plots the results for two subjects: one with intrinsic bias for responding "Category 2" (that is, with a negative criterion location for neutral cues) and one with an intrinsic bias for responding "Category 1" (that is, with a positive criterion location for neutral cues). For the subjects displayed in Fig. 3, the mid-point of the two predictive cues criteria ($c_{PredMid}$) is located in-between the no-bias ($c_{NoBias} = 0$) and the neutral cue ($c_{neutral}$) criteria. Thus, for these two subjects, the predictive cues led to a reduction but not an elimination of the intrinsic response bias (in line with the "reduce" hypothesis).

We tested whether the same overall effect occurred in the whole group of 70 subjects. To do so, we determined the number of subjects for whom the midpoint of the predictive criteria ($c_{PredMid}$) fell (1) on the opposite side of c_{NoBias} compared to $c_{neutral}$, (2) in-between c_{NoBias} and $c_{neutral}$, and (3) beyond $c_{neutral}$ (see Section 2.8 for more details). The number of subjects in each of these categories was 15/36/19 (21.4%/51.4%/27.1%). Then, using these numbers we performed two-sided Binomial tests to check if the midpoint of the predictive criteria ($c_{PredMid}$) was centered either on c_{NoBias} or $c_{neutral}$. We found that for a significant proportion of subjects (78.6%), $c_{PredMid}$ fell on the same side of c_{NoBias} as $c_{neutral}$ (p = 0.000002). Similarly, for a significant proportion of subjects (72.9%), $c_{PredMid}$ fell on the same side of $c_{neutral}$ as c_{NoBias} (p = 0.0002). In other words, $c_{PredMid}$ was centered neither around c_{NoBias} nor around $c_{neutral}$ but tended to fall in-between them.

These binomial tests clearly falsified both the "eliminate" and "maintain" hypotheses and thus strongly supported the "reduce" hypothesis. However, these analyses were limited in two important ways. First, for many subjects, $c_{neutral}$ had a very low absolute value (e.g., 0.1) thus making it likely that estimation error would make $c_{PredMid}$ fall outside of the interval between c_{NoBias} and $c_{neutral}$. Therefore, the Binomial tests may have underestimated the strength of the effects. Second, the Binomial tests do not allow us to determine the degree to which predictive cues reduce the influence of the intrinsic response bias.

To overcome these limitations, we performed regression analyses that quantified how the intrinsic bias, $c_{neutral}$, can be used to predict the midpoint of the predictive criteria, $c_{PredMid}$. Such regression analyses are less likely to be biased by the presence of small intrinsic biases for some of the subjects and can quantify more precisely how much the intrinsic bias is reduced in the presence of predictive cues. Specifically, the β value from this regression can be used as a measure of the degree to which the response bias is maintained in the presence of predictive cues (the "eliminate" and "maintain" hypotheses predict β values of 0 and 1, respectively, whereas the "reduce" hypothesis predicts a β value between 0 and 1).

Across our 70 subjects, we found $\beta = 0.64$, suggesting that predictive cues reduced the influence of intrinsic bias by about a third (Fig. 4). This β value was both significantly greater than 0 (F (1,68) = 120.29, $p = 1.1 \times 10^{-16}$) and significantly smaller than 1 (F (1,68) = 38.04, $p = 4.3 \times 10^{-8}$), thus confirming that both the "eliminate" and "maintain" hypotheses can be rejected with a high degree of certainty. One subject appeared to have a particularly strong intrinsic bias, $c_{neutral}$, and thus be a potential outlier. Removing this subject only slightly increased the estimated slope ($\beta = 0.69$), which remained significantly greater than 0 and smaller than 1 (both p's < 0.00003). Further, we checked for subjects with extreme biases in individual conditions. First, we looked for subjects who responded "Category 1" in any condition either less than 5% of the time or more than 95% of the time and found one such subject. Using the even more lenient limits of 10% and 90% resulted in three such subjects with extreme biases. According to both definitions of extreme bias, excluding subjects with such biases did not affect our conclusions that both the "eliminate" and "maintain" hypotheses have to be rejected (all p's < 0.000004).

We further performed the same regression analyses for each of Experiments 1–3 separately to ensure that the above results were not due to the aggregation of different experiments. All experiments resulted in β values between 0 and 1 (β = 0.6, 0.87, and 0.33 in Experiments 1–3, respectively). These β values were significantly greater than 0 for all three experiments (Experiment 1: F (1,28) = 88.75, $p = 3.5 \times 10^{-10}$; Experiment 2: F(1,19) = 31.55,

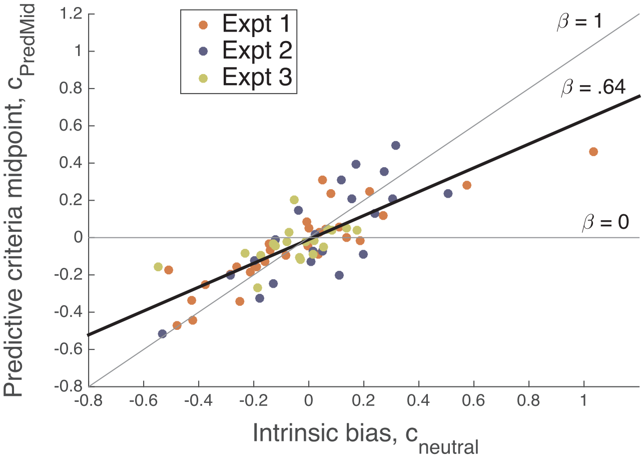

Fig. 4. Regression analyses results. The intrinsic bias, $c_{neutral}$, was used to predict the midpoint of the predictive criteria, $c_{PredMid}$. The obtained β value of 0.64 was significantly different than both 0 (falsifying the "eliminate" hypothesis, which implies that $c_{neutral}$ has no relationship with $c_{PredMid}$) and 1 (falsifying "maintain" hypothesis, which implies that $c_{neutral} = c_{PredMid}$). Each dot represents one subject and colors indicate the experiment in which the subject participated. The thick black line indicates the line of best fit, whereas the thin gray lines indicate the predictions of the "eliminate" and "maintain" hypotheses.

$p = 0.00002$; Experiment 3: F(1,17) = 6.9, $p = 0.018$) and significantly smaller than 1 for two of the three experiments (Experiment 1: F (1,28) = 39.08, $p = 9.3 \times 10^{-7}$; Experiment 2: F(1,19) = 0.73, $p = 0.4$; Experiment 3: F(1,17) = 28.39, $p = 0.00006$). These results confirm that the "reduce" hypothesis that predictive cues reduce but do not eliminate intrinsic bias is supported in individual experiments too.

To further establish the robustness of our conclusions, we explored whether the same effects can be observed with an alternative way of estimating bias. In these control analyses, we quantified bias as simply the proportion of "Category 1" responses (P(resp = Cat1), see Section 2.7). We then computed this alternative bias measure for each cue type and estimated the equivalent quantities to our SDT analyses. Specifically, we tested the relationship between intrinsic bias, $P_{neutral}$(resp = Cat1), and the midpoint of the biases associated with the two predictive cues, $P_{PredMid}$(resp = Cat1).

We found very similar results to our SDT-based analyses. Specifically, for a significant proportion of subjects (78.6%) $P_{PredMid}$(resp = Cat1) fell on the same side of P_{NoBias}(resp = Cat1) as $P_{neutral}$(resp = Cat1) ($p = 0.000002$). Similarly, for a significant proportion of subjects (64.3%) $P_{PredMid}$(resp = Cat1) fell on the same side of $P_{neutral}$(resp = Cat1) as P_{NoBias}(resp = Cat1) ($p = 0.02$). Further, a regression analysis across our 70 subjects produced $\beta = 0.54$, suggesting that predictive cues reduced the influence of intrinsic bias by about half (Fig. 5). This β value was both significantly greater than 0 (F (1,68) = 71.24, $p = 3.5 \times 10^{-12}$) and significantly smaller than 1 (F (1,68) = 52.84, $p = 4.6 \times 10^{-10}$). Finally, the individual β values for each experiment again fell between 0 and 1 (β = 0.56, 0.75, and 0.1 in Experiments 1–3, respectively). These β values were significantly greater than 0 for two of the three experiments (Experiment 1: F (1,28) = 48.59, $p = 1.4 \times 10^{-7}$; Experiment 2: F(1,19) = 29.12, $p = 0.00003$; Experiment 3: F(1,17) = 0.86, $p = 0.37$) and significantly smaller than 1 for two of the three experiments (Experiment 1: F (1,28) = 29.53, $p = 8.5 \times 10^{-6}$; Experiment 2: F(1,19) = 3.17, $p = 0.09$; Experiment 3: F(1,17) = 77.81, $p = 9.4 \times 10^{-8}$). Thus, our results are largely insensitive to the exact bias measure used.

One potential confound in all analyses so far is that in two of our three experiments, we had substantially more total trials with predictive cues compared to neutral cues. This could be problematic since it means that we would have a smaller estimation error when computing subjects' bias for predictive compared to neutral cues. If the estimation error is large and subjects have minimal true bias, then the reduction in bias for predictive cues may in fact be purely the result of smaller estimation error which brings the estimated bias closer to the location of no bias. To ensure that our results are not due to such statistical artefact, we re-did all analyses when considering only the first $\frac{N_{neutral}}{2}$ trials from each predictive cue type (i.e., cues for Category 1 and cues for Category 2) where $N_{neutral}$ is number of neutral-cue trials for a given subject. These re-analyses thus ensured that the same number of trials were used in computing the bias for neutral and predictive cues. We found that all previously significant results remained significant in the re-analyses and the effect sizes showed very little change. For example, in the critical regression analyses, we obtained a β value of 0.62 for the SDT analyses (the β value was 0.64 when all trials were considered) and 0.54 for the alternative bias measure (the β value was again 0.54 when all trials were considered). Similar results were also obtained if we considered last, rather than the first, $\frac{N_{neutral}}{2}$ trials from each predictive cue type. Therefore, our results cannot be explained as a statistical artefact of the different number of trials in the predictive and neutral cue conditions.

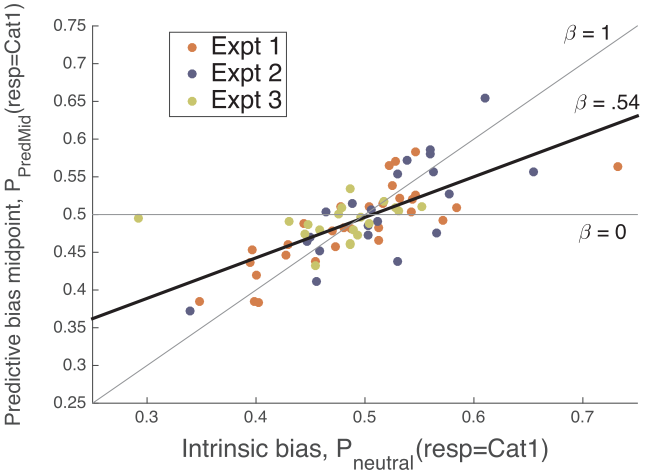

Fig. 5. Regression analyses for an alternative measure of bias. We quantified bias as the probability of giving a "Category 1" response rather than as the SDT criterion and again observed a similar relationship between intrinsic bias, $P_{neutral}$(resp = Cat1), and the midpoint of the predictive criteria, $P_{PredMid}$(resp = Cat1). As in the SDT analyses, the obtained β value of 0.54 was significantly different than both 0 and 1. Each dot represents one subject and colors indicate the experiment in which the subject participated. The thick black line indicates the line of best fit, whereas the thin gray lines indicate the predictions of the "eliminate" and "maintain" hypotheses.

In each analysis above, the data from each subject were reduced to a single scalar value. However, it is important to test whether we can falsify the "eliminate" and "maintain" hypotheses on the level of individual subjects too (Regenwetter & Robinson, 2017). To address this question, we combined the data from the two predictive cues and used Binomial tests for each subject to determine whether the observed proportion of "Category 1" responses across the two predictive conditions was different from 0.5 (thus testing the "eliminate" hypothesis) and the proportion of "Category 1" responses for neutral cues (thus testing the "maintain" hypothesis).

We found that 23 out of our 70 subjects (32.9%) showed significant subject-level bias (i.e., $p < 0.05$ in the Binomial test on their individual data). A Binomial test on the group level demonstrated that this proportion is significantly higher than the 5% of significant subject-level bias expected by chance ($p = 2.1 \times 10^{-13}$). Therefore, these subject-level analyses strongly falsify the "eliminate" hypothesis and show that individual-subject biases remain in the presence of predictive cues. Similarly, 11 of the 70 subjects (15.7%) showed subject-level bias that is significantly different from the bias for neutral cues, which is also significantly higher than the 5% expected by chance ($p = 0.0007$).

Therefore, both the “eliminate” and “maintain” hypotheses are falsified on the level of individual subjects too.

4. Discussion

Understanding perceptual decision making on a mechanistic level requires insight into the nature of subjects’ response biases. Here we investigated how two sources of response bias – intrinsic bias and bias induced by predictive cues – interact with each other. Across three experiments, we observed that predictive cues reduce but do not eliminate intrinsic response bias. These findings demonstrate both the malleability and rigidity of people’s intrinsic biases.

4.1. The mechanisms of bias reduction

Why do predictive cues reduce intrinsic bias? At least four different possibilities exist. We examine each of them below.

First, the reduction in bias could stem from normative considerations, which can typically be expressed within the framework of Bayesian Decision Theory (Maloney & Mamassian, 2009). Within this framework, predictive cues should have a multiplicative influence on the likelihood ratio implied by the criterion location (Rahnev & Denison, 2018). The multiplicative influence on the likelihood ratio is in turn equivalent to an additive influence for the SDT criterion location c (Rahnev, Koizumi, McCurdy, D’Esposito, & Lau, 2015). Therefore, normative considerations would imply that the predictive criteria should be centered around c_{NoBias} or $c_{neutral}$ depending on one’s assumption on whether predictive cues should eliminate or maintain the intrinsic bias. What is important here is that neither scenario would result in a reduction of intrinsic bias based on normative considerations alone. This is not to say that our findings deviate from normativity but simply that additional factors need to be introduced for normativity to make predictions consistent with our results.

Second, it is possible that intrinsic biases are something that people are aware of but are not motivated to overcome. According to this account, the attempt to incorporate the predictive cues into one’s decisions provides extra motivation to curtail the intrinsic bias. However, although motivation could indeed be part of the story here, it does not explain why a bias in a particular direction would arise in the first place.

Third, it is possible that subjects followed one of two possible mixture strategies where with probability P they simply answered according to the cue and with probability 1-P they either responded with exactly the same bias as for neutral cues or responded with no bias at all ($c = 0$). These strategies may be able to fit our data while still postulating that the true underlying bias is either maintained or eliminated. However, both of these strategies result in substantial decreases in d' for predictive cues due to the fact that they involve substantial “criterion jitter” (Rahnev & Denison, 2018), that is, the fact that the criterion changes markedly from trial to trial. To quantify the exact decrease, we fit these two models to the data for each subject (see Supplementary Results) and found that while d' for predictive cues was on average 1.44 in the empirical data, it decreased to 1.27 in the model where bias was maintained ($t(69) = 4.32$, $p = 0.00005$) and to 1.28 in the model where bias was eliminated ($t(69) = 3.84$, $p = 0.0003$). This substantial decrease in d' compared to the empirical data suggests that these mixture strategies cannot explain our data.

Fourth, it could be that the very act of integrating the externally-provided priors into one’s decision making suppresses the factors that gave rise to the intrinsic bias. According to this account, the reduction of bias is an automatic and unconscious process. This is the account that appears most plausible to us at the moment.

This brief discussion reveals that it is currently unclear what the exact mechanisms of the bias reduction are. We have investigated the phenomenon using descriptive methods but many potential mechanisms can give rise to this observed bias reduction, some of which do not postulate a true reduction of internal bias but simply a decision strategy that makes behavior appear less biased (e.g., the third possibility above). Currently, any explanation for the observed effect would necessarily be speculative and will likely remain so until we know where intrinsic biases come from.

4.2. The source of intrinsic bias

What causes the intrinsic biases observed in our experiments? Perhaps the most trivial reason for observing response biases is the limited number of trials used. Indeed, for subjects with a completely unbiased criterion of exactly 0, the measured criterion based on a few hundred trials would be expected to take positive values for some subjects and negative values for others. However, using data from Rahnev, Nee, Riddle, Larson, and D’Esposito (2016), we previously demonstrated that response criteria remained stable over four sessions of the same experiment conducted on separate days (Rahnev & Denison, 2018). Similarly, many other types of stable idiosyncratic biases have been described in the literature (Finlayson et al., 2017; García-Pérez & Alcalá-Quintana, 2011; Kosovicheva & Whitney, 2017; Wexler et al., 2015). In the current experiment, we found that intrinsic bias as measured in one condition (in the presence of neutral cues) persists in a completely different condition (in the presence of predictive cues). This persistence further establishes the existence of stable individual differences in response bias. Taken together, these considerations strongly suggest that the observed intrinsic response bias is not simply caused by noise in the estimation procedure.

If intrinsic bias is not an experimental artifact, it appears that it must be the product of heuristic algorithms of decision making. Indeed, it is widely recognized that humans have limited resources and thus often use heuristic computations (Gershman, Horvitz, & Tenenbaum, 2015; Gigerenzer & Selten, 2002). Nevertheless, it remains unclear what type of heuristic computations would give rise to response biases and why such heuristics are adopted in the first place. Our current findings that the intrinsic bias is reduced but not eliminated in the presence of predictive cues provides one of the few empirical phenomena related to response bias that can begin to unveil its source. The finding that predictive cues lead to a reduction of intrinsic bias of about one third to a half adds additional quantitative precision that can be exploited in creating and testing computational models that postulate potential sources of intrinsic bias. We suggest that the source of subjects’ intrinsic bias should be recognized as one of the most important questions in the pursuit of a mechanistic understanding of perceptual decision making.

4.3. Relationship to other types of bias

In this paper, we use the term “bias” to signify an overall predisposition to choose one of two stimulus categories. However, many other types of biases have been described that alter responses on a trial-by-trial basis. For example, perceptual responses have been shown to be influenced by the cost to act (Hagura, Haggard, & Diedrichsen, 2017), previous stimuli (Cicchini, Mikellidou, & Burr, 2017; Fischer & Whitney, 2014), and previous decisions (Abrahamyan, Silva, Dakin, Carandini, & Gardner, 2016; Fritsche, Mostert, & de Lange, 2017; Manassi, Liberman, Kosovicheva, Zhang, & Whitney, 2018). Such biases pull responses in different directions on different trials and interact with biases that remain stable across the whole experiment (Finlayson et al., 2017; García-Pérez & Alcalá-Quintana, 2011; Kosovicheva & Whitney, 2017; Linares et al., 2019; Wexler et al., 2015). It remains an open question how these different types of biases trade off against each other.

4.4. Novel predictions

We expect that the reduction in intrinsic bias for predictive cues observed in current experiments would generalize across different domains. A particularly intriguing generalization concerns serial dependence (Fischer & Whitney, 2014; Treisman & Faulkner, 1984; Yu & Cohen, 2009). One common type of serial dependence is the tendency to repeat the same response in two-choice tasks. A particular response on one trial can therefore be considered to act as a predictive cue for the next trial. We would therefore predict that serial dependence also acts to reduce the intrinsic response bias. Testing this prediction would require that one includes sufficient number of "neutral" trials. In the context of serial dependence, a "neutral" trial is a trial that is not preceded by a response toward either stimulus category. Practically, neutral trials can be trials presented immediately after a distractor task or after a short break. Future experiments should address whether the results obtained here indeed generalize to serial dependence.

Declaration of Competing Interest

None.

Acknowledgements

We thank Medha Shekhar and Rachel Denison for helpful comments on a previous version of this paper. This work was supported by the National Institute Of Mental Health of the National Institutes of Health under Award Number R56MH119189 to D.R. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

D. Rahnev developed the study concept. M. Hu and D. Rahnev performed the data analysis and interpretation. D. Rahnev drafted the manuscript and M. Hu provided critical revisions. All authors approved the final version of the manuscript for submission.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2019.06.016.

References

Abrahamyan, A., Silva, L. L., Dakin, S. C., Carandini, M., & Gardner, J. L. (2016). Adaptable history biases in human perceptual decisions. *Proceedings of the National Academy of Sciences, 113*(25), E3548–E3557. https://doi.org/10.1073/pnas.1518786113.

Ackermann, J. F., & Landy, M. S. (2015). Suboptimal decision criteria are predicted by subjectively weighted probabilities and rewards. *Attention, Perception & Psychophysics, 77*(2), 638–658. https://doi.org/10.3758/s13414-014-0779-z.

Bang, J. W., & Rahnev, D. (2017). Stimulus expectation alters decision criterion but not sensory signal in perceptual decision making. *Scientific Reports, 7*(1), 17072. https://doi.org/10.1038/s41598-017-16885-2.

Bohil, C. J., & Maddox, W. T. (2001). Category discriminability, base-rate, and payoff effects in perceptual categorization. *Perception & Psychophysics, 63*(2), 361–376.

Bohil, C. J., & Maddox, W. T. (2003). A test of the optimal classifier's independence assumption in perceptual categorization. *Perception & Psychophysics, 65*(3), 478–493.

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General, 121*(2), 177–194. https://doi.org/10.1037/0096-3445.121.2.177.

Cheadle, S., Egner, T., Wyart, V., Wu, C., & Summerfield, C. (2015). Feature expectation heightens visual sensitivity during fine orientation discrimination. e14 *Journal of Vision, 15*(14), https://doi.org/10.1167/15.14.14.

Cicchini, G. M., Mikellidou, K., & Burr, D. (2017). Serial dependencies act directly on perception. *Journal of Vision, 17*(14), 6. https://doi.org/10.1167/17.14.6.

de Lange, F. P., Rahnev, D., Donner, T. H., & Lau, H. (2013). Prestimulus oscillatory activity over motor cortex reflects perceptual expectations. *The Journal of Neuroscience, 33*(4), 1400–1410. https://doi.org/10.1523/JNEUROSCI.1094-12.2013.

Finlayson, N. J., Papageorgiou, A., & Schwarzkopf, D. S. (2017). A new method for mapping perceptual biases across visual space. *Journal of Vision, 17*(9), 5. https://doi.org/10.1167/17.9.5.doi.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience, 17*(5), 738–743. https://doi.org/10.1038/nn.3689.

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.

García-Pérez, M. A., & Alcalá-Quintana, R. (2011). Interval bias in 2AFC detection tasks: Sorting out the artifacts. *Attention, Perception, & Psychophysics, 73*(7), 2332–2352. https://doi.org/10.3758/s13414-011-0167-x.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273–278. https://doi.org/10.1126/science.aac6076.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality*. Cambridge, MA: MIT Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley & Sons Ltd.

Hagura, N., Haggard, P., & Diedrichsen, J. (2017). Perceptual decisions are biased by the cost to act. e18422 *ELife, 6*. https://doi.org/10.7554/eLife. 18422.

Hanks, T. D., & Summerfield, C. (2017). Perceptual decision making in rodents, monkeys, and humans. *Neuron, 93*(1), 15–31. https://doi.org/10.1016/j.neuron.2016.12.003.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory, 7*(5), 344–354. https://doi.org/10.1037/0278-7393.7.5.344.

Jiang, J., Summerfield, C., & Egner, T. (2013). Attention sharpens the distinction between expected and unexpected percepts in the visual brain. *Journal of Neuroscience, 33*(47), 18438–18447. https://doi.org/10.1523/JNEUROSCI.3308-13.2013.

Kok, P., Mostert, P., & de Lange, F. P. (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.

Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H., & De Lange, F. P. (2012). Attention reverses the effect of prediction in silencing sensory signals. *Cerebral Cortex, 22*(9), 2197–2206. https://doi.org/10.1093/cercor/bhr310.

Kosovicheva, A., & Whitney, D. (2017). Stable individual signatures in object localization. *Current Biology, 27*(14), R700–R701. https://doi.org/10.1016/j.cub.2017.06.001.

Linares, D., Aguilar-Lleyda, D., & López-Moliner, J. (2019). Decoupling sensory from decisional choice biases in perceptual decision making. *ELife, 8*. https://doi.org/10.7554/eLife.43994.

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior, 78*(3), 567–595. https://doi.org/10.1901/jeab.2002.78-567.

Maddox, W. T., & Bohil, C. J. (1998). Base-rate and payoff effects in multidimensional perceptual categorization. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 24*(6), 1459–1482.

Maddox, W. T., & Dodd, J. L. (2001). On the relation between base-rate and cost-benefit learning in simulated medical diagnosis. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 27*(6), 1367–1384.

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience, 26*(1), 147–155. https://doi.org/10.1017/S0952523808080905.

Manassi, M., Liberman, A., Kosovicheva, A., Zhang, K., & Whitney, D. (2018). Serial dependence in position occurs at the time of perception. *Psychonomic Bulletin & Review, 25*(6), 2245–2253. https://doi.org/10.3758/s13423-018-1454-5.

Morales, J., Solovey, G., Maniscalco, B., Rahnev, D., de Lange, F. P., & Lau, H. (2015). Low attention impairs optimal incorporation of prior knowledge in perceptual decisions. *Attention, Perception & Psychophysics, 77*(6), 2021–2036. https://doi.org/10.3758/s13414-015-0897-2.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences, 41*(e223), 1–66. https://doi.org/10.1017/S0140525X18000936.

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence leak in perceptual decision making. *Psychological Science, 26*(11), 1664–1680. https://doi.org/10.1177/0956797615595037.

Rahnev, D., Lau, H., & De Lange, F. P. (2011). Prior expectation modulates the interaction between sensory and prefrontal regions in the human brain. *Journal of Neuroscience, 31*(29), 10741–10748.

Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S., & D'Esposito, M. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *Proceedings of the National Academy of Sciences, 113*(20), 6059–6064. https://doi.org/10.1073/pnas.1522551113.

Regenwetter, M., & Robinson, M. M. (2017). The construct–behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review, 124*(5), 533–550. https://doi.org/10.1037/rev0000067.

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.

Todorovic, A., & de Lange, F. P. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *The Journal of Neuroscience, 32*(39), 13389–13395. https://doi.org/10.1523/JNEUROSCI.2227-12.2012.

Treisman, M., & Faulkner, A. (1984). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance, 10*(1), 119–139 Retrieved from http://discovery.ucl.ac.uk/20033/.

Wexler, M., Duyck, M., & Mamassian, P. (2015). Persistent states in vision break universality and time invariance. *Proceedings of the National Academy of Sciences, 112*(48), 14990–14995. https://doi.org/10.1073/pnas.1508847112.

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? *Advances in Neural Information Processing Systems, 21*, 1873–1880.