

Reviewed Preprint

v1 • January 28, 2025

Not revised

Reviewed Preprint

v2 • December 4, 2025

Revised by authors

For correspondence:

raccoon65.y@nycu.edu.tw

swwu@nycu.edu.tw

Competing interests:

No competing interests declared

Reviewing editor:

Andreea Oliviana Diaconescu, University of Toronto, Canada

© 2025, Wang et al. This article is

distributed under the terms of the

[Creative Commons Attribution](#)

[License](#), which permits unrestricted

use and redistribution provided that

the original author and source are

credited.

Detecting Regime Shifts: Neurocomputational Substrates for Over- and Underreactions to Change

Mu-Chen Wang¹✉, George Wu², Shih-Wei Wu^{1,3}✉

¹Institute of Neuroscience, National Yang Ming Chiao Tung University, Taipei, Taiwan • ²Booth School of Business, University of Chicago, Chicago, United States • ³Brain Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan

eLife Assessment

This study offers **valuable** insights into how humans detect and adapt to regime shifts, highlighting distinct contributions of the frontoparietal network and ventromedial prefrontal cortex to sensitivity to signal diagnosticity and transition probabilities. The combination of an innovative task design, behavioral modeling, and model-based fMRI analyses provides a **solid** foundation for the conclusions; however, the neuroimaging results have several limitations, particularly a potential confound between the posterior probability of a switch and the passage of time that may not be fully controlled by including trial number as a regressor. The control experiments intended to address this issue also appear conceptually inconsistent and, at the behavioral level, while informing participants of conditional probabilities rather than requiring learning is theoretically elegant, such information is difficult to apply accurately, as shown by well-documented challenges with conditional reasoning and base-rate neglect. Expressing these probabilities as natural frequencies rather than percentages may have improved comprehension. Overall, the study advances understanding of belief updating under uncertainty but would benefit from more intuitive probabilistic framing and stronger control of temporal confounds in future work.

<https://doi.org/10.7554/eLife.104684.2.sa3>

Abstract

The world constantly changes, with the underlying state of the world shifting from one regime to another. The ability to detect a regime shift, such as the onset of a pandemic or the end of a recession, significantly impacts individual decisions as well as governmental policies. However, determining whether a regime has changed is usually not obvious, as signals are noisy and reflective of the volatility of the environment. We designed an fMRI paradigm that examines a stylized regime-shift detection task. Human participants showed systematic over- and underreaction: Overreaction was most commonly seen when signals were noisy but when environments were stable and change is possible but unlikely. By contrast, underreaction was observed when signals were precise but when environments were unstable and hence change was more likely. These behavioral signatures are consistent with the *system-neglect* computational hypothesis, which posits that sensitivity or lack thereof to system parameters (noise and volatility) is central to these behavioral biases. Guided by this computational framework, we found that individual subjects' sensitivity to system parameters were represented by two distinct brain networks. Whereas a frontoparietal network selectively represented individuals' sensitivity to signal noise but not environment volatility, the ventromedial prefrontal cortex (vmPFC) showed the opposite pattern. Further, these two networks were involved in different aspects of regime-shift computations: while vmPFC correlated with subjects' beliefs about change, the frontoparietal network represented the strength of evidence in favor of regime shifts. Together, these results

suggest that regime-shift detection recruits belief-updating and evidence-evaluation networks and that under- and overreactions arise from how sensitive these networks are to the system parameters.

Significance Statement

Judging whether the world has changed, from the onset of a market boom to the end of a pandemic, is ubiquitous. The ability to detect regime shifts not only impacts individual decisions but also governmental policies. However, these judgments are hard to make because the signals we receive are noisy and reflective of the volatility of the environment. We find that people overreact to changes when they receive noisy signals in stable environments, but underreact when facing precise signals in unstable environments. Under- and overreactions can be read out by distinct brain networks according to their sensitivity in responding to different environmental parameters that impact regime changes. This suggests that parameter selectivity at the network level guides regime-shift detection.

Introduction

Judging whether the world has changed is ubiquitous, from public health officials grappling with whether a pandemic surge has peaked, central banks figuring out whether inflation is easing, investors discerning whether the electric car market is getting traction, or romantic partners divining whether a relationship has soured. In all of these examples, individuals must update their beliefs that the world has changed based on a noisy signal, such as a drop in positive pandemic cases or a romantic partner's suddenly mysterious behavior. In some cases, epidemiological or statistical models provide guidance. However, in many, if not most cases, the determination of whether a regime shift has occurred is made intuitively (Sanders & Manrodt, 2003 [↗](#)).

We investigate intuitive judgments of regime-shift detection using a simple empirical paradigm (Massey & Wu, 2005 [↗](#); Seifert et al., 2023 [↗](#)). Although this paradigm abstracts away some complications of real-world change detection, it maintains the most central features of the problem: normatively, regime-shift judgments reflect the *signals* from the environment as well as knowledge about the *system* that produces the signals. The most recent time series of inflation rates, pandemic cases, and sales of electric cars are all examples of signals. When pandemic cases continue to decline in recent weeks, one might infer a shift from pandemic to non-pandemic regime, only to learn a few weeks later that pandemic has resurged. Indeed, signals such as the latest pandemic cases are seldom precise indications of the true state of the world. Put differently, signals are, by and large, noisy. The noisier the signals are, the less *diagnostic* they are of the underlying regime.

In addition, signals are affected by how likely the regime shifts from one to another (transition probability). These two fundamental features or *system parameters*—the diagnosticity of the signals and transition probability—can be conceptualized as two independent aspects of the system that generates the signals. Previous works on regime-shift detection has found that people tend to overreact to change when they receive noisy signals (low signal diagnosticity) but nonetheless are in a stable environment (small transition probability). By contrast, precise signals (high signal diagnosticity) in an unstable environment (large transition probability) typically results in underreaction (Benjamin, 2019 [↗](#); Brown & Steyvers, 2009 [↗](#); Massey & Wu, 2005 [↗](#)).

Massey and Wu (2005) [↗](#) proposed that over- and underreactions reflect *system neglect*—the tendency to respond primarily to signals and secondarily to the system parameters that produces the signals. The system-neglect hypothesis was derived from theoretical accounts of the determinants of confidence by Griffin and Tversky (1992) [↗](#). To explain system neglect, consider someone who is making judgments on whether a stock market had shifted from the bear to the bull market regime and has been given information about recent stock returns (signals), how frequent regime shifts happen (transition probability), and how similar the two regimes are (signal diagnosticity). If her judgments are solely based on the signals and not affected by transition probability and signal diagnosticity, she shows a *complete* neglect of the system

parameters. Broadly, system neglect describes a lack of sensitivity—compared with normative Bayesian updating—to the system parameters. In the case of regime-shift detection, this leads to insufficient belief revision (i.e., underreaction) in diagnostic and unstable environments, where Bayesian updating requires a larger change in beliefs, and excessive belief change in noisy and stable environments (i.e., overreaction), where Bayesian updating calls for less pronounced belief revision.

Empirical patterns akin to system neglect is not only observed in regime-shift detection, but also in other domains such as confidence judgments (Griffin & Tversky, 1992 [↗](#); Kraemer & Weber, 2004 [↗](#)), demand forecasting (Kremer et al., 2011 [↗](#)), and pricing decisions (Seifert et al., 2023 [↗](#)). Under- and overreactions have been an active research topic in financial economics, often measured as reactions to stock market changes or firm news (Baker & Wurgler, 2007 [↗](#); Barberis et al., 1998 [↗](#); Daniel et al., 1998 [↗](#); De Bondt & Thaler, 1985 [↗](#); Nelson et al., 2001 [↗](#)). At the neurobiological level, change detection has been investigated in the context of reinforcement learning in dynamic environments where changes in the state of the world, such as reward distributions, take place during the experiments (Soltani & Izquierdo, 2019 [↗](#)). Different behavioral paradigms, most notably reversal learning, and computational models were developed to investigate its neurocomputational substrates (Behrens et al., 2007 [↗](#); Izquierdo et al., 2017 [↗](#); McGuire et al., 2014 [↗](#); Muller et al., 2019 [↗](#); Nassar et al., 2010 [↗](#); Payzan-LeNestour & Bossaerts, 2011 [↗](#); Payzan-LeNestour et al., 2013 [↗](#)). Key findings on the neural implementations for such learning include identifying brain areas and networks that track volatility in the environment (rate of change) (Behrens et al., 2007 [↗](#)), the uncertainty or entropy of the current state of the environment (Muller et al., 2019 [↗](#)), participants' beliefs about change (Kao et al., 2020 [↗](#); McGuire et al., 2014 [↗](#); Payzan-LeNestour & Bossaerts, 2011 [↗](#)), and their uncertainty about whether a change had occurred (Kao et al., 2020 [↗](#); McGuire et al., 2014 [↗](#)). Evidence from several of the aforementioned studies (Behrens et al., 2007 [↗](#); Kao et al., 2020 [↗](#); McGuire et al., 2014 [↗](#)) suggests that the dorsomedial frontal cortex (DMFC) is critical to learning in dynamic environments, as information about volatility, subjective beliefs and uncertainty about change converge in this brain region.

But how do biases in change detection arise in the brain? Although reinforcement learning studies provide valuable insights into change detection in the learning process, it remains unclear how biases in change detection—under- and overreactions to change—arise at the neural algorithmic and implementation levels. For example, it is unclear how a certain brain area, such as DMFC, that had been shown to represent environmental volatility, would contribute to under- and overreactions to change. In order to systematically characterize under- and overreactions, it would be critical to (1) adopt a well-established behavioral paradigm that robustly elicits these behavioral phenomena and (2) have computational frameworks suitable for developing neural hypotheses regarding under- and overreactions. To address these issues, in this study, we adopted the regime-shift detection task from Massey and Wu (2005 [↗](#)) and their system-neglect computational framework. At the behavioral level, the regime-shift task is a well-established paradigm that robustly elicits under- and overreactions to change. At the algorithmic and implementation levels, the system-neglect framework provides a straightforward neurocomputational hypothesis regarding under- and overreactions. It predicts that for brain areas involved in regime-shift detection, under- and overreactions arise from their sensitivity or lack thereof in response to the system parameters.

We replicated previous behavioral findings on under- and overreactions (Massey & Wu, 2005 [↗](#)). Using blood-oxygen-level-dependent (BOLD) functional magnetic resonance imaging (fMRI), we reported three key findings. First, we identified two distinct brain networks involved in regime-shift detection, with the ventromedial prefrontal cortex (vmPFC) and ventral striatum in representing subjects' reported beliefs about change and a frontoparietal network in evaluating the strength of change evidence. Second, we found that these two networks selectively respond to different system parameters: while the frontoparietal network represents individual subjects' sensitivity to signal diagnosticity but not transition probability, the vmPFC shows the opposite pattern. Third, the neural sensitivity profiles were signal-dependent: frontoparietal network only

represented individuals' sensitivity to signal diagnosticity when signals consistent with change appeared. By contrast, vmPFC represented individuals' sensitivity to transition probability regardless of whether subjects received signals consistent or inconsistent with change. Such signal-dependent representations led us to further examine and subsequently verify that they are indeed key properties of our system-neglect computational model. Together, these results suggest that regime-shift detection is implemented jointly by a belief-updating network (vmPFC-striatum) and evidence evaluation network (frontoparietal network) and that their sensitivity in response to different environmental parameters contribute to under- and overreactions to change. More broadly, we showed that neural data can reveal important properties of computational models that are overlooked in theoretical treatments and behavioral analyses.

Results

In our *regime-shift detection* task (Fig. 1A), in each trial, subjects saw a series of sequentially presented sensory signals (red or blue balls). They were told that the signals came from one of two regimes, the red regime or the blue regime (Fig. 1B). Regimes were symmetric, for example, with a red regime consisting of 60 red balls and 40 blue balls and the corresponding blue regime consisting of 60 blue balls and 40 red balls. Each trial started with the red regime but could shift to the blue regime before each of the 10 periods in a trial. After seeing a new signal in each period, subjects provided a probability estimate that the current regime was the blue regime, i.e., a posterior probability of a regime shift. They were also instructed that once the regime has shifted from the red to the blue during a trial, the regime would remain in the blue regime until the end of the trial, i.e., the blue regime was a trapping or absorbing state. Our experimental paradigm hence follows Massey and Wu (2005). Note that, during a trial, subjects did not receive feedback—after making probability estimates in each period—on whether the regime had shifted and the monetary bonus earned as a result of accuracy in probability estimates (see *Methods* for details). Hence, subjects had no access to information about accuracy and rewards as she or he was making probability estimates.

We manipulated two system parameters, transition probability and signal diagnosticity (Fig. 1C). Transition probability, q , with possible values are 0.01, 0.05, and 0.1, specified the probability that the regime would shift from the red to the blue regime in any period. Signal diagnosticity, d , with possible values of 1.5, 3, and 9, captured the degree to which the two regimes differed. For example, an environment with high signal diagnosticity (e.g., $d = 9$) indicated that there were 9 times more red balls than blue balls in the red regime (a 90: 10 Red to Blue ratio) and 9 times more blue balls than the red balls in the blue regime (a 90: 10 Blue to Red ratio). Therefore, the weight that a signal (blue or red ball) carried was captured by the signal diagnosticity: in a low diagnostic environment, a blue signal most likely reflects no change in regime ($d = 1.5$, example on the left of Fig. 1D). By contrast, in a highly diagnostic environment, a blue signal very likely reveals a shift in regime ($d = 9$, example on the right of Fig. 1D). At the beginning of each trial, subjects were informed about the transition probability and signal diagnosticity in that trial. In the example trial sequence (Fig. 1A), the transition probability (indicated by “switch probability” in Fig. 1A) is 0.1 while the signal diagnosticity (indicated by “color ratio” in Fig. 1A) is 1.5, with the red regime consisting of 60 red balls and 40 blue balls and the blue regime consisting of 40 red balls and 60 blue balls.

To establish the neural representations for regime-shift estimation, we performed three fMRI experiments ($n = 30$ subjects for each experiment, 90 subjects in total). Experiment 1 was the main experiment, while Experiments 2 to 3 were control experiments that ruled out two important confounds (Fig. 1E). Experiment 2 was designed to establish whether the results from Experiment 1 was truly about regime-shift probability estimation or simply reflected probability estimation. In Experiment 1, the subjects had to estimate the probability that the current regime is the blue regime. Experiment 2 was identical to Experiment 1 except that the environments were stationary (no transition from one regime to another was possible) in Experiment 2 as in Edwards (1968) classic “bookbag-and-poker chip” studies. Subjects in both experiments had to estimate the probability that the current regime is the blue regime, but this estimation corresponded to the

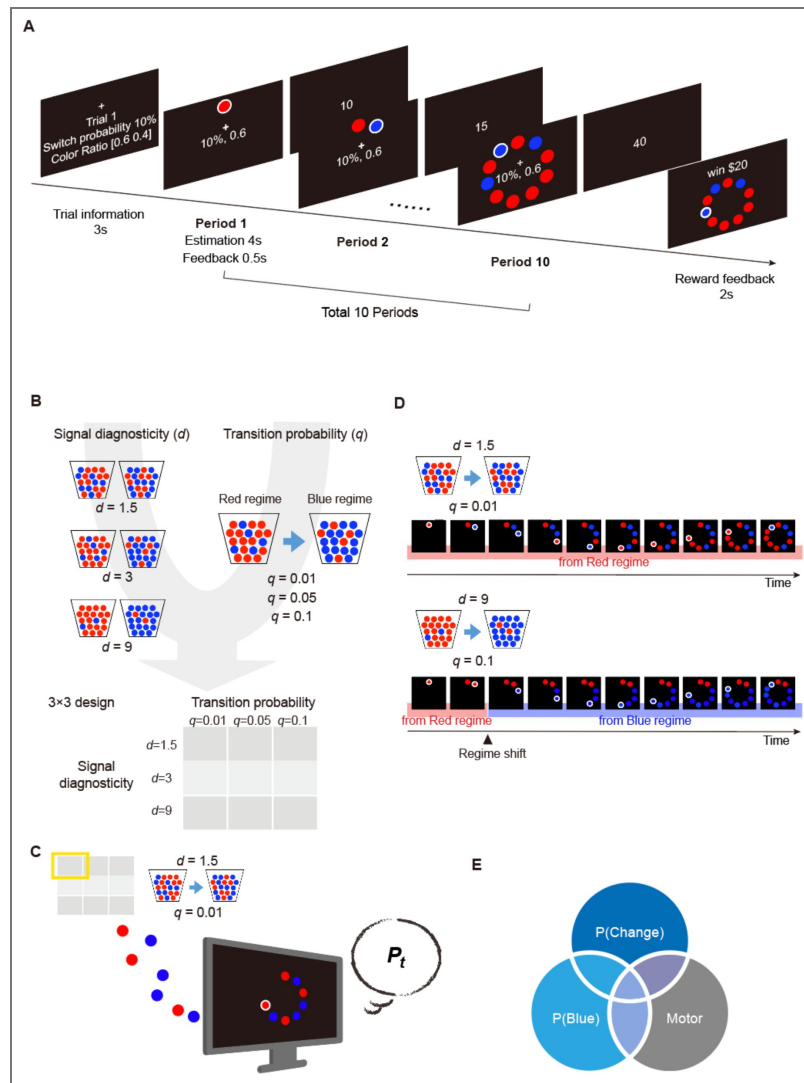


Figure 1. The regime-shift detection task.

A. Trial sequence. In each trial, the subjects saw a sequence of red and/or blue signals and were told that these signals were drawn from one of the two regimes, a Red regime and a Blue regime. Both regimes were described as urns containing red and blue balls. The Red regime contained more red balls, while the Blue regime contained more blue balls. Each trial always started at the Red regime but could shift to the Blue regime in any of the 10 periods according to some transition probability (q). At the beginning of a trial, information about transition probability (shown as “switch” probability in the illustration) and signal diagnosticity (shown as “color ratio”) were revealed to the subjects. In this example, the transition probability is 0.1 and signal diagnosticity is 1.5. See main text for more detailed descriptions. **B.** Manipulation of the system parameters, i.e., transition probability (q) and signal diagnosticity (d). We independently manipulated the q (3 levels) and d (3 levels), resulting in a 3x3 factorial design. **C.** An example of a particular combination of the system parameters from the 3x3 design. Here the system that produces the signals has a $q = 0.01$ transition probability and $d = 1.5$ signal diagnosticity. Signals were sequentially presented to subjects. After each new signal appeared (a period), subjects provided a probability estimate (P_t) of a regime shift. **D.** Two example trials sequences. The example on the left shows the sequence of 10 periods of blue and red signals where $d = 1.5$ and $q = 0.01$. In this example, the regime was never shifted. The example on the right shows the sequence of periods where $d = 9$ and $q = 0.1$. In this example, the regime was shifted from the Red to the Blue regime in Period 3 such that the signals shown starting at this period were drawn from the Blue regime. **E.** We performed three fMRI experiments (30 subjects in each experiment) to investigate the neural basis of regime-shift judgments. Experiment 1 was the main experiment looking at regime shift—which corresponds to $P(\text{Change})$ in the Venn diagram—while Experiments 2 and 3 were the control experiments that ruled out additional confounds. In both Experiments 1 and 2, the subjects had to estimate the probability that signals came from the blue regime. But unlike Experiment 1, in Experiment 2, which corresponds to $P(\text{Blue})$, no regime shift was possible. In Experiment 3, the subjects were simply asked to enter a number with a button-press setup identical to Experiments 1 and 2. Therefore, Experiment 3 (Motor) allowed to rule out motor confounds.

estimates of regime change only in Experiment 1. Therefore, activity that correlated with probability estimates in Experiment 1 but not in Experiment 2 can be uniquely attributed to representing the probability estimates of regime shift. Experiment 3 was specifically designed to address the motor confounds—whether neural signals correlated with probability estimates arose from motor preparation and/or execution, as the subjects made button presses to enter their probability estimates. By comparing the fMRI results of these experiments, we were therefore able to establish the neural representations that can be uniquely attributed to the probability estimates of regime-shift.

Behavioral evidence for over- and underreactions to change

Our analyses used subjects' probability estimates of a regime shift, P_t , for each period, $t = 1, \dots, 10$. We found that subjects were in general responsive to the system parameters, with higher P_t when the transition probability was larger. We also found that subjects tended to give more extreme P_t under high signal diagnosticity than low diagnosticity (Fig. S1 in Supplementary Information, SI). In addition, we used a measure of belief revision, $\Delta P_t = P_t - P_{t-1}$. In Fig. 2A we show examples of P_t and ΔP_t from a subject. On the left, the subject was in a stable environment (small transition probability, $q = 0.01$) and faced two regimes that were very similar to each other (low signal diagnosticity, $d = 1.5$). The red and blue signals (10 periods) were what the subject encountered during a trial. On the right, the subject was in an unstable environment ($q = 0.1$) and faced two regimes that were very different ($d = 9$).

To examine over- and underreactions to change, we compared subjects' belief revision, $\Delta P_t = P_t - P_{t-1}$, $t = 2, \dots, 10$, with belief revision predicted by the Bayesian model, $\Delta P_t^B = P_t^B - P_{t-1}^B$ (see Fig. 2A for illustrations). ΔP and ΔP_t^B respectively capture how much subjects and a normative Bayesian change probability estimates in response to a new signal. When $\Delta P_t > \Delta P_t^B$, it indicates larger belief revision than the normative Bayesian, i.e., an overreaction. By contrast, $\Delta P_t < \Delta P_t^B$ indicates smaller belief revision, i.e., an underreaction. We therefore use $IO = \Delta P_t - \Delta P_t^B$ as an Index of Overreaction (IO). We found that subjects tended to overreact to change ($IO > 0$) when they received noisy signals (i.e., low signal diagnosticity, $d = 1.5$) and when the environment was stable (small transition probability, $q = 0.01$). By contrast, underreaction ($IO < 0$) was most commonly observed when they were in unstable environments (large transition probability, $q = 0.1$) and with clear signals (i.e., high signal diagnosticity, $d = 9$) (Fig. 2B). These patterns of over- and underreactions were consistent with findings in (Massey & Wu, 2005) and the system-neglect hypothesis, which posits a tendency to respond primarily to the signals and secondarily to the system that generates the signals (Massey & Wu, 2005; Seifert et al., 2023). According to the system-neglect hypothesis, responding secondarily to the system is synonymous with a lack of sensitivity to the system parameters, which leads to underreactions in unstable environments with precise signals, and overreactions in stable environments with noisy signals.

Following Massey and Wu (2005), we quantitatively model these belief revisions using the system-neglect model (see *System-neglect model* in *Methods*). The model is a parameterized version of the normative Bayesian model that allows for distortion of the system parameters via weighting parameters for transition probability (α) and signal diagnosticity (β). In short, α reflects distortion of transition probability, with $\alpha \times q$ in the system-neglect model capturing a decision maker's effective transition probability (q). For example, if $\alpha = 4$ when $q = 0.01$, the decision maker effectively treats a 0.01 transition probability as if it were 0.04. By contrast, β captures the extent to which the decision maker overweighs or underweighs signal diagnosticity (d^β) when faced with a signal. For example, if $\beta = 2$ when $d = 1.5$, subjects would treat a blue signal by updating the odds ratio for change by 1.5^2 , or 2.25 rather than 1.5.

In the system-neglect model, we estimated the weighting parameters separately for each level of transition probability and signal diagnosticity, i.e., $\alpha_i \times q_i$ and $d_j^{\beta_j}$, where α_1 , α_2 , and α_3 correspond respectively to transition probabilities of 0.01, 0.05, and 0.1, and β_1 , β_2 , and β_3

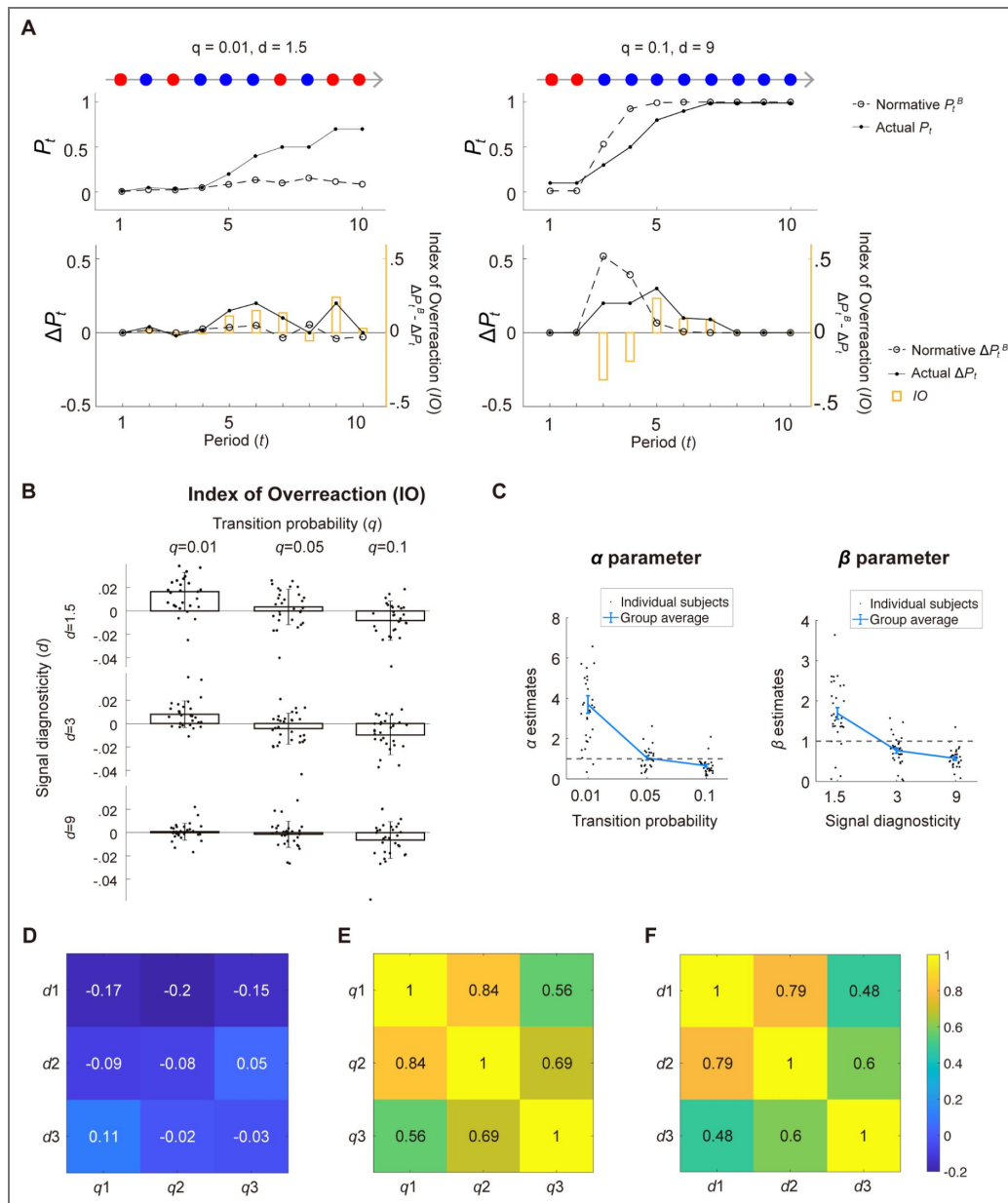


Figure 2. Behavioral results for Experiment 1.

A. Illustrations of Over- and underreactions. Left column: stable environment ($q = 0.01$) with noisy signals ($d = 1.5$) and the 10 periods of red and blue signals a subject encountered. Right column: unstable environment ($q = 0.1$) with precise signals ($d = 9$). Top row: we plot a subject's actual probability estimates (P_t , solid line) and the normative Bayesian posterior probability (P_t^B , dashed line). Bottom row: belief revision shown by the subject ($\Delta P_t = P_t - P_{t-1}$, solid line) and the Bayesian belief revision (ΔP_t^B , dashed line). The orange bars represent $\Delta P_t - \Delta P_t^B$, which we define as the Index of Overreaction (IO; vertical axis in orange on the right). **B.** Over- and underreactions to change. The mean IO (across all 30 subjects) is plotted as a function of transition probability and signal diagnosticity. Subjects overreacted to change if $IO > 0$ and underreacted if $IO < 0$. Error bars represent ± 1 standard error of the mean. **C.** Parameter estimates of the system-neglect model. Left graph: Weighting parameter (α) for transition probability. Right graph: Weighting parameter (β) for signal diagnosticity. Dashed lines indicate parameter values equal to 1, which is required for Bayesian updating. **D-F.** Sensitivity to transition probability and signal diagnosticity are independent. **D.** Correlation between α and β estimates at different levels of transition probability (q_1 to q_3) and signal diagnosticity (d_1 to d_3). All pairwise Pearson correlation coefficients (indicated by the values on the table that were also color coded) were not significantly different from 0 ($p > .05$). **E.** Pearson correlation coefficients of α estimates between different levels of transition probability. All pairwise correlations were significantly different from 0 ($p < .05$). **F.** Pearson correlation coefficients of β estimates between different levels of signal diagnosticity. All pairwise correlations were significantly different from 0 ($p < .05$).

correspond respectively to signal diagnosticity of 1.5, 3, and 9. In contrast to the Bayesian model which implies $\alpha_i = \beta_j = 1$ for all i, j , the system-neglect model requires that that $\alpha_i > \alpha_{i+1}$ and $\beta_j > \beta_{j+1}$ because it would effectively capture a lack of sensitivity to the system parameters.

We fit the model to P_t for each subject separately and found parameter estimates consistent with system neglect (Fig. 2C). The mean estimates of α were 3.69, 1.04 and 0.65 respectively for $q = 0.01, 0.05$ and 0.10 . The parameters indicated that, on average, when $q = 0.01$, subjects treated as if it were 0.0369. By contrast, when $q = 0.10$, the subjects treated it as if it were 0.065. Thus, a factor of 10 in actual transition probability (0.01 vs. 0.1) was reduced to a factor of less than 2 (0.0369 vs. 0.065) in effective transition probability. For signal diagnosticity, the mean parameter estimates of β were 1.69, 0.77 and 0.57 for $d = 1.5, 3$, and 9 , respectively. Thus, subjects updated their beliefs $1.5^{1.69} = 1.98$ when $d_1 = 1.5$ and $9^{0.57} = 3.50$ when $d_3 = 9$. Normatively, the change in odds ratio between the two conditions should have been $d_3/d_1 = 6$ but, consistent with system neglect, was considerably smaller, $3.50/1.98 = 1.76$. Together, large parameter estimates ($\alpha > 1, \beta > 1$) at low signal diagnosticity (noisy signals) and low transition probability (stable environments) capture overreactions to changes, while small parameter estimates ($\alpha < 1, \beta < 1$) at large signal diagnosticity (precise signals) and large transition probability (unstable environments) reflect underreactions to change. These results replicate the findings by Massey and Wu (2005), with the pattern of over- and underreactions as predicted by the system neglect hypothesis. Critically, the degree of system neglect can be captured by the negative trend of the parameter estimate as a function of the system parameter levels (Fig. 2C): the steeper the slope, the larger the system neglect. We found a similar pattern on β in Experiment 2 (one of the control experiments) where environments were stationary (no transition probability) and signal diagnosticity was manipulated (Fig. S2 in SI) (Benjamin, 2019; Tversky et al., 1990).

We performed a parameter recovery analysis to examine whether the fitting procedure gave reasonable parameter estimates (Wilson & Collins, 2019). First, we simulated each subject's probability estimation data based on the system-neglect model by using that subject's parameter estimates. Second, we fitted the system-neglect model to the simulated data. Third, we computed the correlation across subjects between the estimated parameters and the parameter values we used to simulate data. Fourth, we repeated the above steps by adding independent white noise to the simulated data. Across different levels of noise, we found good parameter recovery (Pearson's r for transition probability $r \geq 0.9533$ across different noise levels, Pearson's r for signal diagnosticity $r \geq 0.9515$ across different noise levels) (Figs. S3 and S4 in SI). In addition, we found that the empirical results (Fig. 2C) can be reproduced by the system-neglect model (Figs. S5 and S6 in SI). That is, we used each subject's parameter estimates to compute the period-wise probability estimates according to the system-neglect model and used these probability estimates to compute and plot index of overreaction (IO). The patterns of IO based on the system-neglect model (Fig. S6 in SI) were very similar to those based on subjects' actual data (Fig. 2B).

We next examined whether the way subjects respond to different system parameters is similar. It is possible, for example, that subjects who showed stronger (or weaker) distortion of transition probability (captured by α parameter) also showed stronger (or weaker) distortion of signal diagnosticity (captured by β parameter). There was no significant correlation between α and β parameters (Fig. 2D). However, we did find within-parameter correlation: subjects who had a higher α_i for a transition probability level i also tended to have a higher α_{i1} for a second transition probability level i' (Fig. 2E), with the same pattern also holding for signal diagnosticity (Fig. 2F). Together, these results suggested that the way an individual decision maker responds to information about the probability of change in the environment (transition probability) has little to do with how she or he responds to information about the similarity between different regimes (signal diagnosticity). But individuals are consistent in responding to a particular system parameter (transition probability or signal diagnosticity) across different levels of the parameter.

fMRI results

We focus our fMRI analyses on addressing three questions. First, what are the brain regions that correlated with subjects' probability estimates of change and belief revision? Second, what are the neural representations for the computational variables contributing to these probability estimates? Third, how might neural responses in the identified brain areas be associated with under- and overreactions to change?

Ventromedial prefrontal cortex and ventral striatum represent probability estimates and belief revision on regime shift

Our first analysis is aimed at identifying brain regions that represented our subjects' regime-shift estimation. To address this question, we used two behavioral measures, namely the period-by-period probability estimates of regime shift, P_t , and the change in P_t between successive periods, ΔP_t . P_t can be regarded as the subjects' posterior probability estimates of regime shift, whereas ΔP_t captures the change in belief (belief revision) about regime shift in the presence of a new signal (see Fig. 3A [for an example on \$P_t\$ and \$\Delta P_t\$](#)).

For P_t , we found that the ventromedial prefrontal cortex (vmPFC) and ventral striatum correlated with this behavioral measure of subjects' belief about change. In addition, many other brain regions, including the motor cortex, central opercular cortex, insula, occipital cortex, and the cerebellum also significantly correlated with P_t (Fig. 3B [see GLM-1 in Methods, and Tables S1 to S3 in SI](#), respectively, for significant clusters of activation using Gaussian random field theory, permutation test on threshold-free-cluster-enhancement statistic, and permutations test on cluster-extent statistic).

For ΔP_t , we also found that the vmPFC and ventral striatum were associated with regime shift belief revision (Fig. 3C [see GLM-1 in Methods and Tables S1 to S3 in SI](#) for significant clusters of activation). While many brain regions correlated with regime-shift probability estimates (P_t), only the vmPFC and ventral striatum also correlated with belief revisions, ΔP_t . We examined the robustness of P_t representations in these two regions with different transformations of P_t : a GLM that replaced P_t with the log odds of P_t , $\ln(P_t/(1 - P_t))$ (Fig. S7 in SI), a GLM (GLM-2 in Methods) that, in addition to P_t , included various task-related variables contributing to P_t as regressors (Fig. S8 in SI), and a GLM that examined P_t separately on periods when change-consistent (blue balls) and change-inconsistent (red balls) signals appeared (Fig. S9 in SI). Each of these analyses showed the same pattern of correlations between P_t and activation in vmPFC and ventral striatum, further establishing the robustness of the P_t findings.

Brain regions shown to correlate with regime-shift probability estimates, P_t , could be driven by motor response because larger estimates predominantly involved right-hand finger presses (see *Methods* for details). To rule out motor confounds, we conducted two control experiments (Experiments 2 and 3) and performed two analyses. First, we examined the neural correlates of probability estimates (P_t in GLM-1) in the control experiments (Experiments 2 and 3). Second, we compared the effect of P_t (GLM-1) between the main experiment (Experiment 1) and the control experiments. In the first analysis, we found that in both control experiments, vmPFC and ventral striatum did not significantly correlate with probability estimates P_t at the whole-brain level (in Experiment 2, no significant clusters of activation at the whole-brain level; see Table S4 for Experiment 3 in SI). In the second analysis, we found that for both vmPFC and ventral striatum, the regression coefficient of P_t was significantly different between Experiment 1 and Experiment 2 (Fig. 3C [and between Experiment 1 and Experiment 3 \(Fig. 3D \[also see Tables S5 and S6 in SI\]\(#\)](#)). In a separate, independent (leave-one-subject-out, LOSO) ROI analysis on vmPFC and ventral striatum, we also found the same results (Fig. S10 in SI). Finally, we note that in GLM-1, we implemented an "action-handedness" regressor to directly address the motor-confound issue, that higher probability estimates preferentially involved right-handed responses for entering higher digits. The action-handedness regressor was parametric, coding -1 if both finger presses involved the left hand (e.g., a subject pressed "23" as her probability estimate when seeing a signal), 0 if using one left finger and one right finger (e.g., "75"), and 1 if both finger presses involved the right

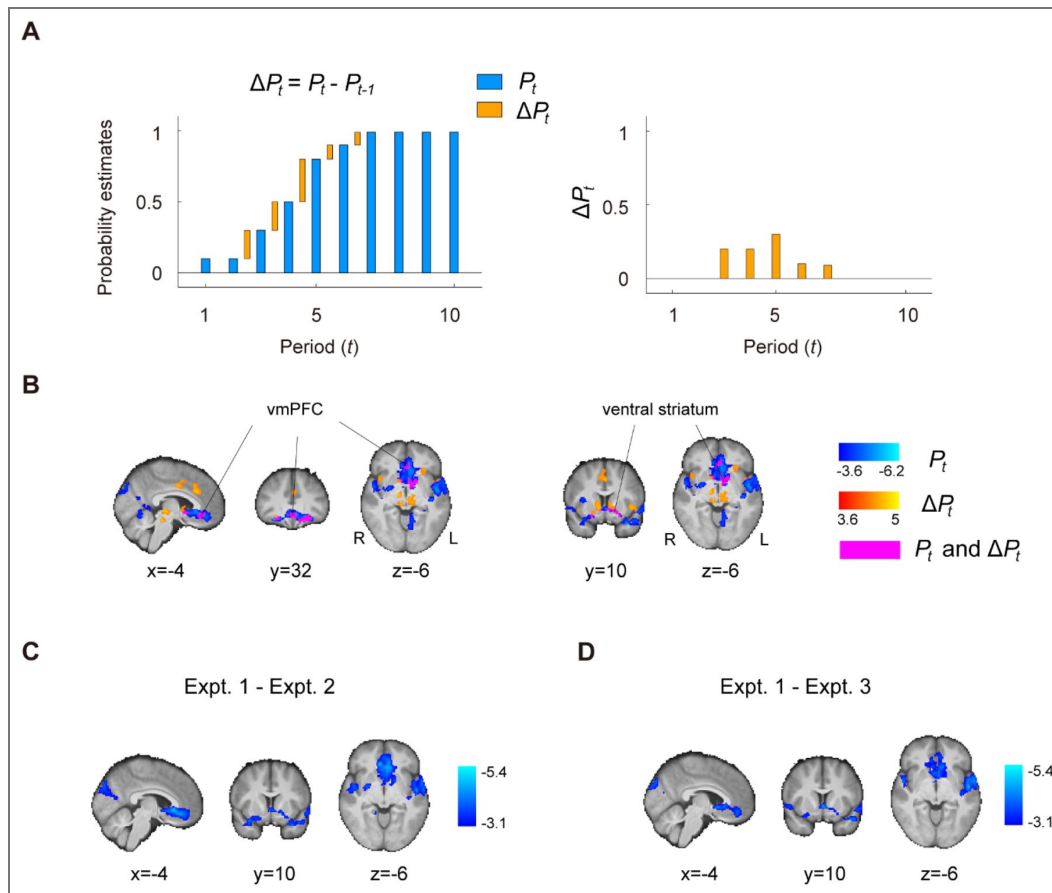


Figure 3. Neural representations for the updating of beliefs about regime shift.

A. An example. Updating is captured by the difference in probability estimates between two adjacent periods (ΔP_t). The Blue bars reflect the period probability estimates (P_t), while yellow bars depict ΔP_t . **B.** Whole-brain results on the main experiment (Experiment 1) showing significant brain regions that correlate with the updating of beliefs about change (ΔP_t). The clusters in orange represent activity correlated with ΔP_t . The blue clusters represent activity correlated with the regime-shift probability estimates (P_t). The magenta clusters represent the overlap between the P_t and ΔP_t clusters. **C-D.** Comparison between experiments. To rule out visual and motor confounds on the P_t results described in **B**, we compared the P_t contrast between the main experiment (Experiment 1) and two control experiments (Experiments 2 and 3). **C.** Whole-brain results on the effect of P_t between Experiments 1 and 2 (Experiment 1 - Experiment 2 on the negative P_t contrast). **D.** Whole-brain results on the effect of P_t between Experiments 1 and 3 (Experiment 1 - Experiment 3 on the negative P_t contrast).

hand (e.g., “90”). Taken together, these results ruled out motor confounds and suggested that vmPFC and ventral striatum represent subjects’ probability estimates of change (regime shifts) and belief revision.

A frontoparietal network represents key variables for estimating regime shifts

Our second analysis is aimed at identifying brain regions that represented key variables contributing to regime-shift estimation. Guided by our theoretical framework and computational models, we focused on two variables, the interaction between signals and signal diagnosticity and intertemporal prior probability of change, and set up a new GLM (GLM-2 in *Methods*) to examine these effects.

Our theoretical framework makes two fundamental predictions. First, a signal should be weighted differently depending on signal diagnosticity, i.e., a blue ball is stronger evidence for change in a highly diagnostic environment (e.g., $d = 9$) than a system in which the red and blue regimes are very similar (e.g., $d = 1.5$). To capture the interaction between signals and signal diagnosticity, we code a blue signal as 1 and a red signal as -1 and multiply the signal code ($s = 1$ or -1) by the natural logarithm of signal diagnosticity, $\ln(d)$ (two examples are shown in Fig. 4A). We term this interaction, $s \times \ln(d)$, the strength of evidence in favor of change or *strength of change evidence* for short. The Bayesian model, as described in *Methods*, critically depends on d^s , computing posterior odds by multiplying prior odds by the likelihood ratio. Thus, the log posterior odds were calculated from both the prior odds and $s \times \ln(d)$. At the whole-brain level, we found that a frontoparietal network including the dorsal medial prefrontal cortex (dmPFC), lateral prefrontal cortex (bilateral inferior frontal gyrus, IFG), and the posterior parietal cortex (bilateral intraparietal sulcus, IPS) represented $s \times \ln(d)$ (Fig. 4A). These brain regions overlap with what is commonly referred to as the frontoparietal control network (Buckner et al., 2013; Seeley et al., 2007; Yeo et al., 2011). Among them, dmPFC sits in the vicinity of dorsomedial frontal cortex (DMFC) shown to represent change probability and uncertainty about change in reinforcement learning (McGuire et al., 2014).

The second prediction our theoretical framework offers concerns the prior probability of a regime shift over time. Specifically, the Bayesian model predicts that the prior probability should increase over time (see two examples in Fig. 4B), with the *intertemporal prior*, in log odds terms, defined as the natural logarithm of the odds in favor of regime shift in the t -th period, $\ln\left(\frac{1-(1-q)^t}{(1-q)^t}\right)$, where q is transition probability and $t = 1, \dots, 10$ is the period (see Eq. 1 in *Methods*). With independent (leave-one-subject-out, LOSO) ROI analysis, we examined whether brain regions in the frontoparietal network (shown to represent strength of change evidence) correlated with intertemporal prior and found that all brain regions, with the exception of dmPFC, in the frontoparietal network correlated with the intertemporal prior (Fig. 4B; dmPFC: $t(29) = -1.69$, $p = 0.10$; left IFG: $t(29) = 2.20$, $p = 0.04$; right IFG: $t(29) = -2.64$, $p = 0.01$; left IPS: $t(29) = -2.35$, $p = 0.03$; right IPS: $t(29) = -2.07$, $p = 0.05$). By contrast, brain regions that represented the intertemporal prior, which we found to be in the right fusiform cortex in the occipitotemporal regions, did not correlate with the strength of change evidence (Fig. S11 in SI).

Finally, we emphasize that these effects—the strength of change evidence and intertemporal prior—cannot be otherwise attributed to probability estimates (P_t) or belief revision (ΔP_t) because both P_t and ΔP_t were included in GLM-2 where these effects were examined. Taken together, these results suggest that the frontoparietal network is critically involved in representing the two key variables for estimating regime shifts, strength of change evidence and intertemporal prior. See Supplementary Tables S7 to S9, respectively, for information about significant clusters of activation using Gaussian random field theory, permutation test on threshold-free-cluster-enhancement (TFCE) statistic, and permutation test on cluster-extent statistic.

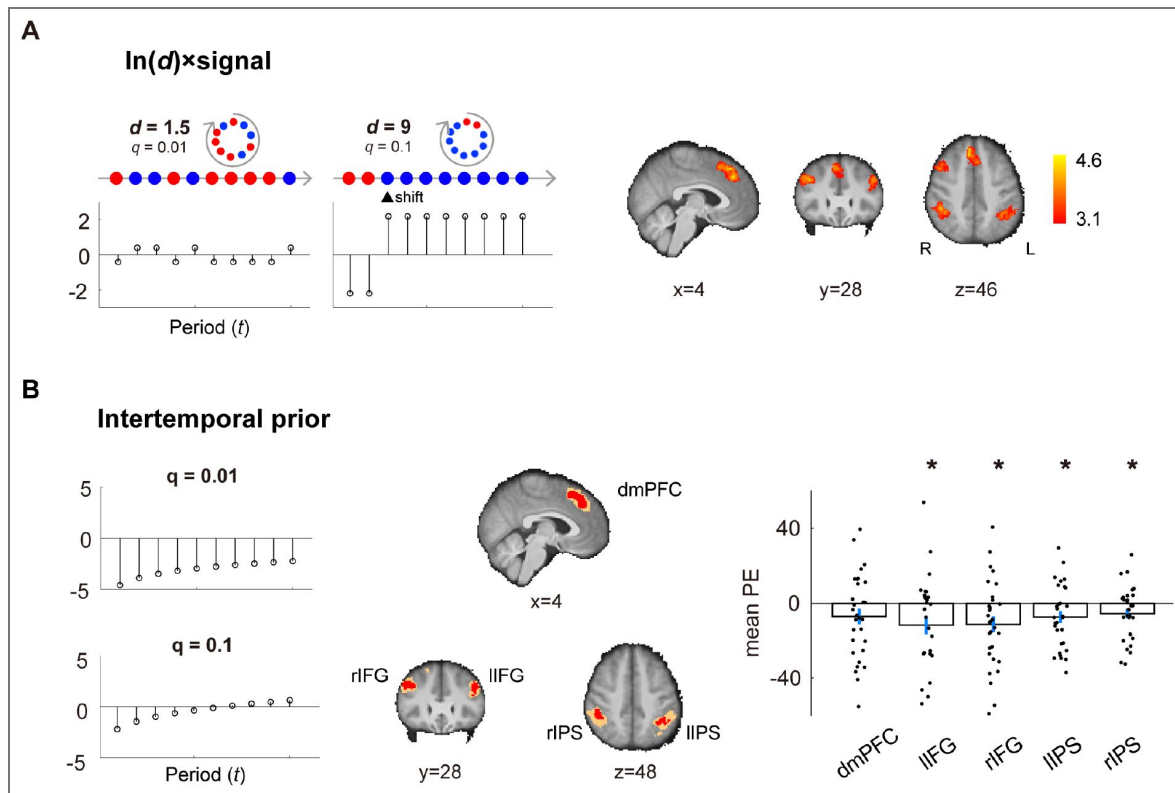


Figure 4. A frontoparietal network represented key variables for regime-shift estimation.

A. Variable 1: strength of evidence in favor of/against regime shifts (strength of change evidence), as measured by the interaction between signal diagnosticity (d) and sensory signal (s), where a blue signal is coded as 1 and a red signal is coded as -1. The x-axis represents the time periods, from the first to the last period, in a trial. The y-axis represents the interaction, $\ln(d) \times s$. Right: whole-brain results showing brain regions in a frontoparietal network that significantly correlated with $\ln(d) \times s$. **B.** Variable 2: intertemporal prior probability of change. Two examples of intertemporal prior are shown on the left graphs. To examine the effect of the intertemporal prior, we performed independent region-of-interest analysis (leave-one-subject-out, LOSO) on the brain regions identified to represent strength of evidence (5A). Due to the LOSO procedure, individual subjects' ROIs (a cluster of contiguous voxels) would be slightly different from one another. To visualize such differences, we used the red color to indicate voxels shared by all individual subjects' ROIs, and orange to indicate voxels by at least one subject's ROI. The ROI analysis examined the regression coefficients (mean PE) of intertemporal prior. The * symbol indicates $p < .05$, ** indicates $p < .01$. dmPFC: dorsomedial prefrontal cortex; IIPS: left intraparietal sulcus; rIPS: right intraparietal sulcus; IIFG: left inferior frontal gyrus; rIFG: right inferior frontal gyrus.

Under- and overreactions are associated with selectivity and sensitivity of neural responses to system parameters

The system-neglect hypothesis posits that under- and overreactions arise from a lack of sensitivity to the system parameters. We can measure individual subjects' sensitivity to system parameters using behavioral data (subjects' probability estimates). Meanwhile, we can also measure sensitivity using neural data. In the following analysis, we examined whether there is a match between the behavioral and neural measures of sensitivity to the system parameters. This would allow us to examine, through the system-neglect framework, whether sensitivity in neural responses to the system parameters are associated with under- and overreactions to change.

We focused on the vmPFC-striatum network and frontoparietal network, as they were shown to be involved in regime-shift detection (Figs. 3 and 4). We examined whether these brain networks show selective preference for a particular system parameter, which we refer to as *parameter selectivity*. We also asked whether parameter selectivity is signal-dependent, i.e., different for signals consistent with change (blue signals) or inconsistent with change (red signals).

We started by defining a behavioral measure of sensitivity to the system parameter. To visualize this measure, we consider two extreme decision makers, a Bayesian and someone who reacts to signals identically across all systems, which we term *complete neglect*. In Fig. 5A (left graph), we use signal diagnosticity (d) to illustrate the pattern of these two decision makers. The vertical axis is $\beta \ln(d)$ and the horizontal axis is the signal-diagnosticity level (d), where β is the weighting parameter on signal diagnosticity in the system-neglect model. A Bayesian (open circles) does not overweight or underweight d , and thus $\beta = 1$. We can then define the *Bayesian slope* by regressing $\beta_i \ln(d_i)$ against $\ln(d_i)$. In this formulation, the Bayesian slope is 1 and it reflects the sensitivity of a Bayesian decision maker to signal diagnosticity. On the other hand, a complete-neglect decision maker is unresponsive to signal diagnosticity, i.e., $\beta_1 \ln(d_1) = \beta_2 \ln(d_2) = \beta_3 \ln(d_3)$. Hence, the *complete-neglect slope* should be 0. These two slopes, the Bayesian slope and the complete-neglect slope, provide the boundaries for system neglect. For each subject, we computed $\beta_i \ln(d_i)$ at each d_i level, where β_i is the estimate for diagnosticity d_i fitted to the system-neglect model (see β_i in Fig. 2B). We then estimated each subject's *behavioral slope* (to distinguish it from the *neural slope* reported later) and use it as a behavioral measure of sensitivity to signal diagnosticity.

For each subject, we estimated two behavioral slopes, one for d , the signal diagnosticity (top row in Fig. 5A), and the other for q , the transition probability (bottom row in Fig. 5A). The right graphs in Fig. 5A shows the behavioral slope for each of the 30 subjects (top: signal diagnosticity; bottom: transition probability). For signal diagnosticity, 28 out of 30 subjects' behavioral slope were within the boundaries. For transition probability, 27 out of 30 subjects' behavioral slope were within the boundaries (between complete neglect and Bayesian). One subject's (subject 6) behavioral slope for q (transition probability) was approximately 2 and clearly outside the boundaries. This subject's data were excluded for further analysis on q (the right two columns in Fig. 5CD).

We found that, across subjects, system neglect was unique from either Bayesian or complete neglect. Subjects' sensitivity to transition probability, as captured by the behavioral slope in Fig. 5A, deviated significantly from the Bayesian slope (comparing subjects' slope with 1, $t(29) = -10.8$, $p < .01$, two-tailed) and from complete neglect slope (comparing subjects' slope with 0, $t(29) = 4.8$, $p < .01$, two-tailed). For signal diagnosticity, subjects' sensitivity to signal diagnosticity was also significantly different from both Bayesian ($t(29) = -12.5$, $p < .01$, two-tailed) and complete neglect ($t(29) = 6.1$, $p < .01$, two-tailed).

However, subjects were closer to complete neglect than to the Bayesian. We tested this by examining whether the behavioral slope, N , was significantly greater or smaller than 0.5, the midpoint between complete neglect (slope of 0) and Bayesian (slope of 1). $N - 0.5 > 0$ indicates that subjects' behavior was in closer alignment with Bayesian. By contrast, $N - 0.5 < 0$ implies behavior closer to complete neglect. We found that, for both transition probability and signal diagnosticity, the behavioral slope was closer to complete neglect than to Bayesian (transition probability: $t(29) = -2.97$, $p < .01$; signal diagnosticity: $t(29) = -3.23$, $p < .01$, two-tailed). Together, these results suggested

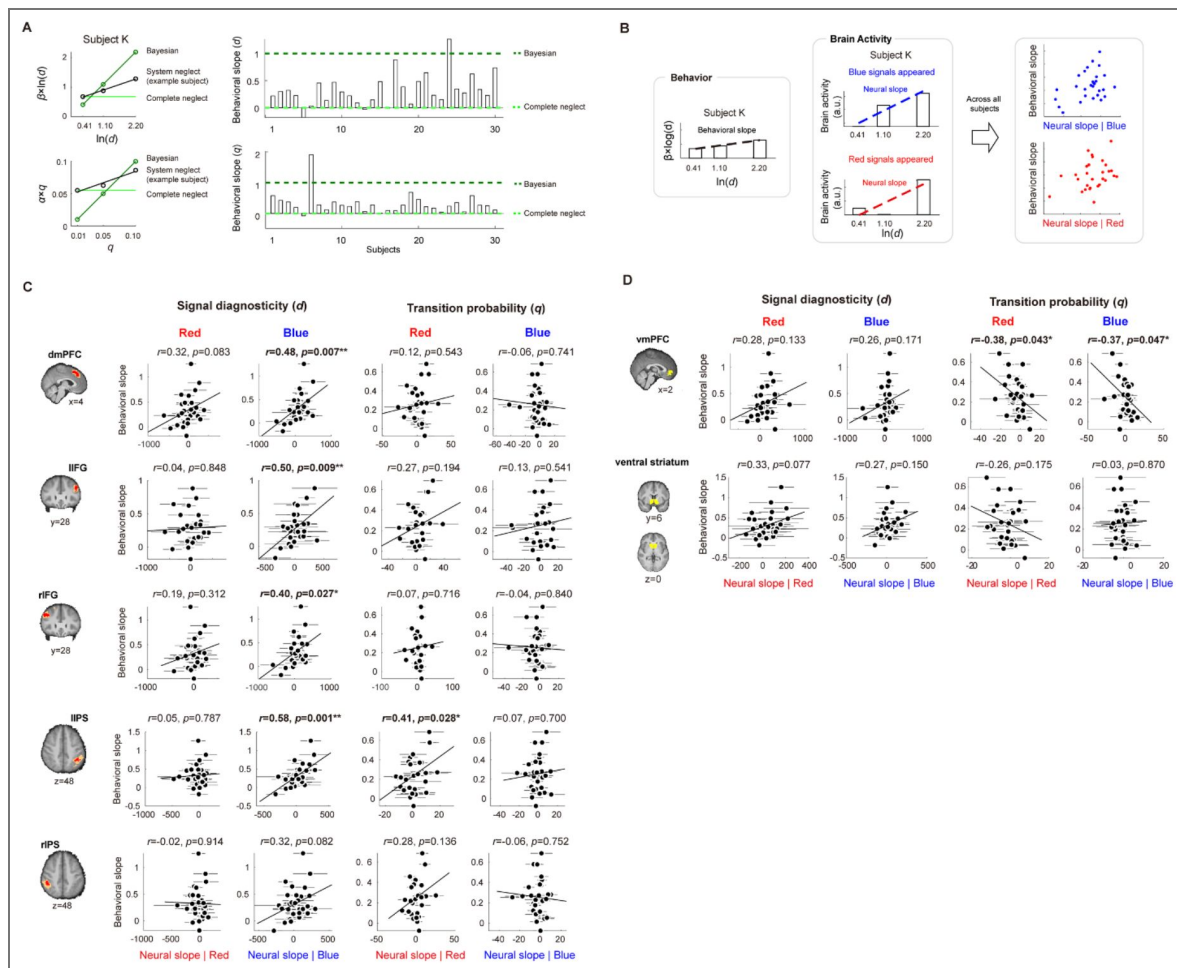


Figure 5. Estimating and comparing neural measures of sensitivity to system parameters with behavioral measures of sensitivity.

A. Behavioral measures of sensitivity to system parameters. For each system parameter, we plot the subjectively weighted system parameter against the system parameter level (top row: signal diagnosticity; bottom row: transition probability). For each subject and each system parameter, we estimated the slope (how the subjectively weighted system parameter changes as a function of the system parameter level) and used it as a behavioral measure of sensitivity to the system parameter (behavioral slope). We also show a Bayesian (no system neglect) decision maker's slope (dark green) and the slope of a decision maker who completely neglects the system parameter (in light green; the slope would be 0). A subject with stronger neglect would have a behavioral slope closer to complete neglect. **B.** Comparison of behavioral and neural measures of sensitivity to the system parameters. To estimate neural sensitivity, for each subject and each system parameter, we regressed neural activity of a ROI against the parameter level and used the slope estimate as a neural measure of sensitivity to that system parameter (neural slope). We also estimated the neural slope separately for blue-signal periods (when the subject saw a blue signal) and red-signal periods. We computed the Pearson correlation coefficient (r) between the behavioral slope and the neural slope and used it to statistically test whether there is a match between the behavioral and neural slopes. **C.** The frontoparietal network selectively represented individuals' sensitivity to signal diagnosticity (left two columns), but not transition probability (right two columns). Further, neural sensitivity to signal diagnosticity (neural slope) correlated with behavioral sensitivity (behavioral slope) only when a signal in favor of potential change (blue) appeared: all the regions except the right IPS showed statistically significant match between the behavioral and neural slopes. By contrast, sensitivity to transition probability was not represented in the frontoparietal network. **D.** The vmPFC selectively represented individuals' sensitivity to transition probability ($r = -0.38$, $p = 0.043$ for red signals; $r = -0.37$, $p = 0.047$ for blue signals), but not signal diagnosticity ($r = 0.28$, $p = 0.13$ for red signals; $r = 0.26$, $p = 0.17$ for blue signals). The ventral striatum did not show selectivity to either transition probability or signal diagnosticity. Error bars represent ± 1 standard error of the mean.

that, while subjects did respond to the system parameters in regime-shift estimation in the correct direction predicted by the Bayesian model, their sensitivity to the system parameters was closer to complete neglect than to normative Bayesian.

For the neural data, we defined a neural measure of sensitivity to the system parameters by estimating how neural responses change as a function of those parameters. Using the signal diagnosticity parameter as an example, for each subject and each ROI separately, we regressed average brain activity at each diagnosticity level against $\ln(d)$. The slope estimate, termed the *neural slope*, from the linear regression gave us a neural measure of sensitivity to signal diagnosticity. To investigate whether the neural sensitivity was signal-dependent, i.e., neural sensitivity in response to the change-consistent signals (blue signals) was different from the change-inconsistent signals (red signals), we separately estimated the neural slope in response to blue and red signals.

After obtaining both the behavioral and neural measures of sensitivity to the system parameters, we then computed the Pearson correlation coefficient between them. We found that the vmPFC-striatum network and frontoparietal network showed clear dissociations in how they corresponded with the system parameters. First, the frontoparietal network represented individual subjects' sensitivity to signal diagnosticity (left two columns in Fig. 5C), but not transition probability (right two columns in Fig. 5C). Notably, patterns of parameter selectivity were remarkably consistent across brain regions in the frontoparietal network: when change-consistent signals (blue signals) appeared, the neural measure of sensitivity from all brain regions in the frontoparietal network except the right IPS significantly correlated with the behavioral measure of sensitivity (second column from the left in Fig. 5B; dmPFC: $r = 0.48$, $p = 0.007$; lIFG: $r = 0.5$, $p = 0.009$; rIFG: $r = 0.4$, $p = 0.027$; lIPS: $r = 0.58$, $p = 0.001$; rIPS: $r = 0.32$, $p = 0.082$). By contrast, when change-inconsistent signals (red signals) appeared, all regions within the frontoparietal network did not significantly correlate with the behavioral measure of sensitivity (first column from the left in Fig. 5B; dmPFC: $r = 0.32$, $p = 0.083$; lIFG: $r = 0.04$, $p = 0.848$; rIFG: $r = 0.19$, $p = 0.312$; lIPS: $r = 0.05$, $p = 0.787$; rIPS: $r = -0.02$, $p = 0.914$). We further tested, for each brain region, whether the difference in correlation was significant using both parametric and nonparametric tests (see *Parametric and nonparametric tests for difference in correlation coefficients* in *Methods*). The results were identical. In the parametric test, we used the Fisher z transformation to transform the difference in correlation coefficients to the z statistic. We found that among the five ROIs in the frontoparietal network, two of them, namely the left IFG and left IPS, the difference in correlation was significant (one-tailed z test; left IFG: $z = 1.8355$, $p = 0.0332$; left IPS: $z = 2.3782$, $p = 0.0087$). For the remaining three ROIs, the difference in correlation was not significant (dmPFC: $z = 0.7594$, $p = 0.2238$; right IFG: $z = 0.9068$, $p = 0.1822$; right IPS: $z = 1.3764$, $p = 0.0843$). We chose one-tailed test because we already know the correlation under change-consistent signals was significantly greater than 0. In the nonparametric test, we performed nonparametric bootstrapping to test for the difference in correlation. We referred to the correlation between neural and behavioral sensitivity at change-consistent (blue) signals as r_{blue} , and that at change-inconsistent (red) signals as r_{red} . Consistent with the parametric tests, we also found that the difference in correlation was significant in left IFG and left IPS (left IFG: $r_{blue} - r_{red} = 0.46$, $p = 0.0496$; left IPS: $r_{blue} - r_{red} = 0.5306$, $p = 0.0041$), but was not significant in dmPFC, right IFG, and right IPS (dmPFC: $r_{blue} - r_{red} = 0.1634$, $p = 0.1919$; right IFG: $r_{blue} - r_{red} = 0.2123$, $p = 0.1681$; right IPS: $r_{blue} - r_{red} = 0.3434$, $p = 0.0631$). In summary, we found that neural sensitivity to signal diagnosticity measured at change-consistent signals significantly correlated with individual subjects' behavioral sensitivity to signal diagnosticity. By contrast, neural sensitivity to signal diagnosticity measured at change-inconsistent signals did not significantly correlate with behavioral sensitivity. The difference in correlation, however, was statistically significant in some (left IPS and left IFG) but not all brain regions within the frontoparietal network.

Second, in contrast to the frontoparietal network, vmPFC in the vmPFC-striatum network showed the opposite pattern of parameter selectivity: vmPFC selectively represented individual subjects' sensitivity to transition probability (right two columns in Fig. 5D), but not to signal diagnosticity (left two columns in Fig. 5D). Selectivity in vmPFC was not signal-dependent: regardless of

change-consistent (blue) or change-inconsistent (red) signals, neural sensitivity to transition probability in vmPFC represented individual subjects' sensitivity to transition probability ($r = -0.38$, $p = 0.043$ for change-inconsistent signals; $r = -0.37$, $p = 0.047$ for change-consistent signals). By contrast, the ventral striatum did not show selectivity to either the transition probability or signal diagnosticity (transition probability: $r = -0.26$, $p = 0.175$ for change-inconsistent signals; $r = 0.03$, $p = 0.870$ for change-consistent signals; signal diagnosticity: $r = 0.33$, $p = 0.077$ for change-inconsistent signals; $r = 0.27$, $p = 0.150$ for change-consistent signals). In summary, these results suggest that vmPFC selectively represented individuals' sensitivity to transition probability, whereas the frontoparietal network selectively represented individuals' sensitivity to signal diagnosticity.

Incorporating signal dependency into system-neglect model led to better models for regime-shift detection

The neural findings on signal dependency (Fig. 5) point to the possibility that participants might respond to the system parameters differently when facing change-consistent and change-inconsistent signals. This led us to ask whether building signal dependency into the system-neglect model would be a better model choice for subjects' behavioral data (probability estimates of regime shift) than the original system-neglect model. To examine this question, we built and fit three new versions of the system-neglect (SN) model (see Table S12 in SI for model-fitting summary) and compared them with the original model (SN-original) (see Table S13 in SI for summary of statistical tests for model comparison). In the signal-dependent β system-neglect model (SN-SigDep- β model), we estimated the β parameters separately at change-consistent and change-inconsistent signals. As a result, in this model there were 6 β parameters—three for change-consistent signals to model each of the three levels of signal diagnosticity and three for change-inconsistent signals—and three α parameters that modeled each of the three levels of transition probability without distinguishing between change-consistent and change-inconsistent signals. In the signal-dependent α system-neglect model (SN-SigDep- α model), we estimated the α parameters separately at change-consistent and change-inconsistent signals. As a result, in this model there were 6 α parameters (three for change-consistent signals and three for change-inconsistent signals) and three β parameters. In the signal-dependent α and β system-neglect model (SN-SigDep- $\alpha\beta$ model), we estimated both α and β parameters separately at change-consistent and change-inconsistent signals (12 total parameters). Compared with SN-original, we found that SN-SigDep- β , SN-SigDep- α , and SN-SigDep- $\alpha\beta$ qualitatively described subjects' behavioral data (belief revision, ΔP_t) better (Fig. 6B-E). Further, we found that estimating α separately at change-consistent and change-inconsistent signals (SN-SigDep- α , Fig. 6D) model improved model fits than estimating β separately (SN-SigDep- β , Fig. 6C), suggesting that subjects responded to transition probability differently when facing change-consistent and change-inconsistent signals more than to signal diagnosticity. Model comparison using Akaike Information Criterion (AIC) revealed that SN-SigDep- $\alpha\beta$ is the best model, followed by SN-SigDep- α , SN-SigDep- β , and SN-original (Fig. 6F). Together, these results suggest that participants showed system-neglect to both transition probability and signal diagnosticity and that they responded to these system parameters differently when facing change-consistent and change-inconsistent signals. In summary, signal dependency in response to system parameters is a new behavioral finding not reported in the original Massey and Wu (2005) study and is largely inspired by the neural sensitivity findings in the current study.

Discussion

We investigated the neurocomputational substrates for regime-shift detection and mechanisms that give rise to under- and overreactions to change. Regime shifts—the transition from one state of the world to another—are present in many daily situations, from the stock market (a change from the bull to the bear market) to the state of a pandemic. Detecting regime shifts can be challenging for at least two reasons. First, the signals we receive from the environments are often noisy. A signal in favor of potential change, for example a drop in pandemic cases, can either inform a true shift in regime or simply reflect noisy fluctuations. Second, the signals we receive

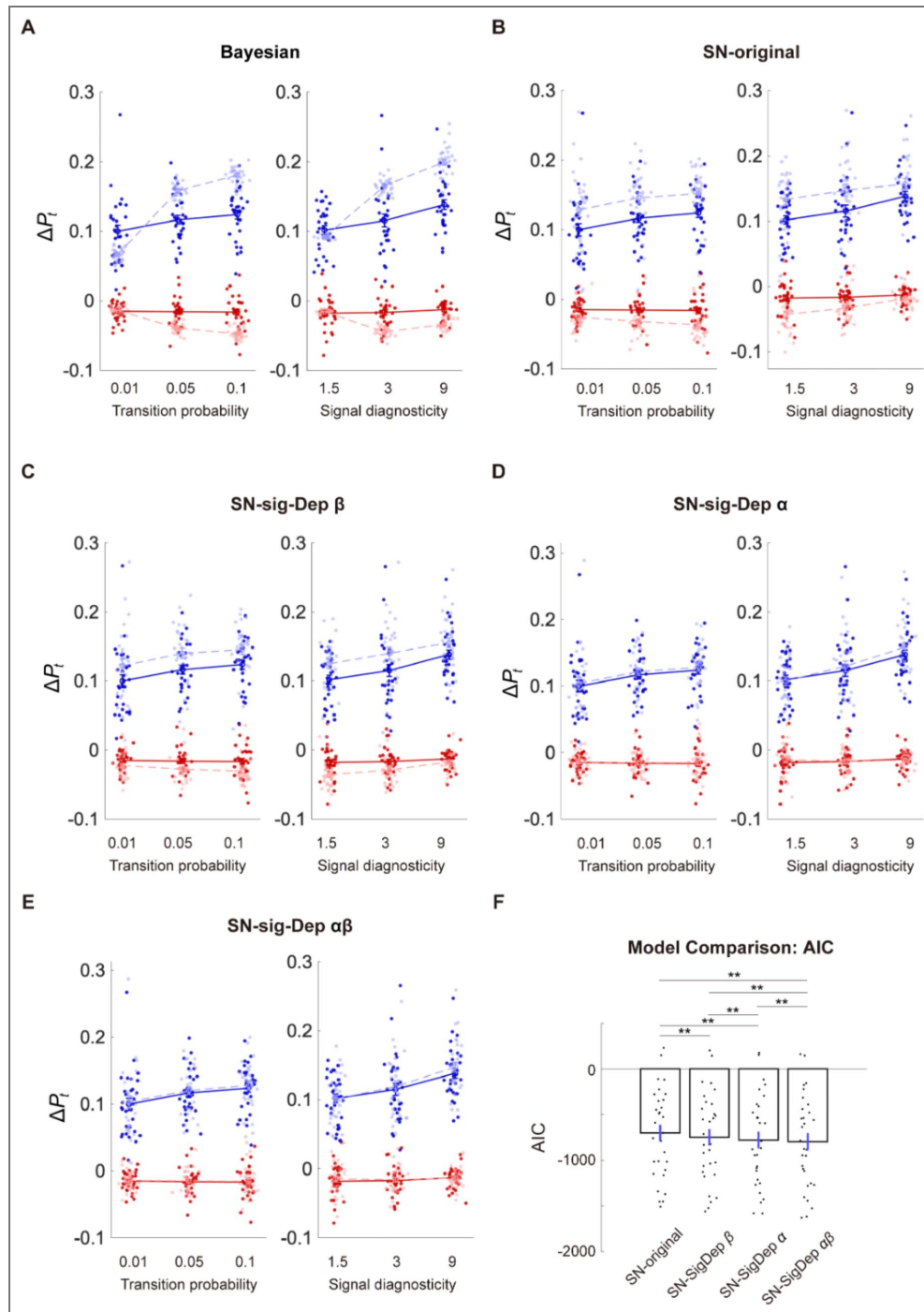


Figure 6. Model comparison.

A-E. Modeling results from five competing models. For each model, we plot subjects' belief revision (ΔP_t) and the model-estimated ΔP_t . Light-colored dots and dashed lines respectively represent the model-estimated ΔP_t at the individual and group levels. Dark-colored dots and solid lines indicate individual subjects' ΔP_t and group-averaged behavioral data respectively. Blue indicates data and model estimates at change-consistent signals; Red indicates data and model estimates at change-inconsistent signals. **A.** Bayesian model. **B.** Original system-neglect model (SN-original). **C.** Signal-dependent β system-neglect model (SN-SigDep- β). **D.** Signal-dependent α system-neglect model (SN-SigDep- α). **E.** Signal-dependent α and β system-neglect model (SN-SigDep- $\alpha\beta$). **F.** Model comparison based on the Akaike Information Criterion (AIC). Lower AIC values indicate better model fits. The bars indicate group mean AIC (averaged across all subjects), while the black dots indicate individual subjects' AIC values. Error bars represent ± 1 standard error of the mean. The * symbol indicates $p < .05$, ** indicates $p < .01$ (paired t-test; see Table S13 in SI for summary of statistical tests).

reflect the volatility of the environment: while some environments are more prone to changes, others are not. To capture these two key features in regime-shift detection, we designed an fMRI task based on Massey and Wu (2005) [\[1\]](#) where subjects made probability judgments about regime shifts and where we manipulated the signal diagnosticity and transition probability. Signal diagnosticity captures the level of noise inherent in the signals, while transition probability reflects the volatility of the environment. Replicating Massey and Wu (2005) [\[1\]](#), we found that overreactions to regime shifts take place when participants received noisy signals (low signal diagnosticity) but when the environments were stable (low transition probability). By contrast, when the signals are more precise but the environments were unstable, participants tended to underreact to changes. These results suggest system neglect— people respond primarily to signals and secondarily to the system that generates the signals (Massey & Wu, 2005 [\[1\]](#)).

At the neurobiological level, we found that regime-shift detection is jointly implemented by two networks, the vmPFC-striatum network and a frontoparietal network. The vmPFC-striatum network represented subjects' probability estimates of change and the revision of probability estimates in the presence of new signals (belief revision). By contrast, the frontoparietal network represented strength of change evidence and intertemporal prior probability of change—two key variables contributing to probability estimation. Guided by the system-neglect framework, we found that under- and overreactions to change are closely associated with the sensitivity of these networks in response to the system parameters—transition probability and signal diagnosticity—that impact regime changes. In particular, the vmPFC represented individual subjects' sensitivity to transition probability, whereas the frontoparietal network represented sensitivity to signal diagnosticity. Together, these findings suggest that selectivity and sensitivity of neural responses to system parameters are key mechanisms that give rise to under- and overreactions.

Our work is closely related to the reversal-learning paradigm—the standard paradigm in neuroscience and psychology to study change detection (Fellows & Farah, 2003 [\[2\]](#); Izquierdo et al., 2017 [\[3\]](#); O'Doherty et al., 2001 [\[4\]](#); Schoenbaum et al., 2000 [\[5\]](#); Walton et al., 2010 [\[6\]](#)). In a typical reversal-learning task, human or animal subjects choose between two options that differ in the reward magnitude or probability of receiving a reward. Through reward feedback the participants gradually learn the reward contingencies associated with the options and have to update knowledge about reward contingencies when contingencies are switched in order to maximize rewards. While a switch in reward contingencies can be regarded as a kind of regime shift, there are three major differences between the reversal-learning paradigm and our regime-shift task. The first difference is about learning. In the reversal-learning paradigm, the subjects must learn both the reward contingencies and the switches through experience. By contrast, in the regime-shift task, the subjects were explicitly informed about the makeup of different regimes and the transition probability. Therefore, participants do not need to learn about the different regimes and the transition probability through experience. The second difference is the kind of behavioral data collected. In our task, we asked the subjects to estimate the probability of change, whereas in the reversal-learning task, the subjects indicate their choice preferences. The third difference is on reward contingency. In the reversal-learning task, a change is specifically about change in reward contingencies, which is not the case in our task.

We believe that these major differences in task design led to three key insights into change detection from our study. The first insight is on over- and underreactions to change. At the behavioral level, we were able to identify situations that led to over- and underreactions. At the theoretical level, we were able to provide a systematic account for these over- and underreactions in the system-neglect hypothesis. Finally, at the neurobiological level, we were able to quantify the degree to which individual subjects neglected the system parameters and use these behavioral measures to unravel the neural mechanisms that give rise to over- and underreactions to change.

The second insight is on the brain networks associated with change detection. In particular, we were able to clarify whether the neural systems involved in change detection in the reversal-learning tasks are contingent on whether rewards are involved. Since the reversal-learning tasks are about learning the reward contingencies and the change in reward contingencies, it would be challenging to infer whether the neural implementations of change detection are dissociable from

reward processing. Indeed, brain regions shown to be involved in the reversal-learning tasks, the OFC, mPFC, striatum, and amygdala, were also found to be highly involved in reward-related learning and value-based decision making. In the current study, unlike in reversal-learning paradigms, regimes were not defined by rewards (e.g., high reward-probability regime vs. low reward-probability regime in reversal learning paradigm). Therefore, estimating the probability of regime shifts in our task did not require considerations for change in reward contingencies. Our findings that vmPFC and ventral striatum represent probability estimates of change and belief revision therefore suggest that these brain regions might be part of a common pathway for change detection in general where changes in the state of the world do not have to be about changes in reward contingencies.

The third insight has to do with the impact of learning on change detection. Under the reversal-learning paradigm, it has been challenging to infer whether there exists a unique change-detection mechanism that is dissociable from reinforcement learning mechanisms. The way to make such inference is through theory, such as implementing a prior for state changes (Bartolo & Averbeck, 2020 [↗](#); Costa et al., 2015 [↗](#)). Unlike the reversal-learning task, participants in our task did not have to learn about the different regimes through experience. Without the confound of reinforcement learning, our results help clarify the roles of change detection on choice behavior by suggesting that independent of learning, there exists a specialized change-detection mechanism in the brain that impacts decision making. This mechanism involves the participation of the vmPFC-striatum network and the frontoparietal network, which partially overlap with the brain regions involved in reversal learning. However, it remains to be seen how learning interacts with change detection. Future investigations can address this question by combining the key features of both the reversal-learning paradigm and regime-shift paradigm.

Outside of the reversal-learning paradigm, previous fMRI studies that investigated learning and belief updating in dynamic environments where change takes place regularly had identified brain regions that represent perceived likelihood of change inferred from participants' choice behavior. Payzan-LeNestour et al. (2013) [↗](#) identified that the posterior cingulate cortex, postcentral gyrus, middle temporal gyrus, hippocampus, and insula correlated with subjects' perceived likelihood of change in a multi-arm bandit task. McGuire et al. (2014) [↗](#) found that subjective change-point probability was represented in a large posterior cluster including occipital, inferior temporal, and posterior parietal cortex. In addition, activity in dorsomedial frontal cortex, posterior cingulate cortex, superior frontal cortex, and anterior insula also positively correlated with change probability. Interestingly, both McGuire et al. (2014) [↗](#) and our results found that the ventral striatum negatively correlated with probability estimates of change. This result suggested that ventral striatum represents probability estimates of change irrespective of whether the task was based on a learning-based paradigm (McGuire et al., 2014 [↗](#)) or non-learning paradigm where information about task-related variables was explicitly revealed to the participants. Further, both McGuire et al. and our results found the involvement of the dorsomedial prefrontal cortex (dmPFC; or dorsomedial frontal cortex in McGuire et al.) in change detection. Our results further suggest that dmPFC is specialized in weighing the strength of change evidence and represents individual subjects' sensitivity to signal diagnosticity, both of which played important roles in contributing to the over- and underreactions to change.

How might our results relate to value-based decision making? In previous studies, vmPFC had been implicated to dynamically track financial risks that carry potential monetary gains or losses. To understand dynamic computations of risk, Schonberg et al. (2012) [↗](#) used a Balloon Analog Risk Task (BART) where subjects decide whether to inflate a simulated balloon through successive pumps for the potential to win larger gains or incur larger losses (if balloon explodes), or to cash out before the balloon explodes. They found that vmPFC activity decreased as subjects pumped and expanded the balloon, suggesting its involvement in estimating the risk of potential losses. Since the explosion of balloon can be regarded as a change in the state of the balloon, as the balloon expands the possibility of such a change in state (regime shift) also increases. In this view, the vmPFC result from Schonberg et al. (2012) [↗](#) was consistent with our finding in that vmPFC

negatively correlated with probability estimates of regime shift. Together, these results add to the existing literature by suggesting that vmPFC is involved in estimating and updating the state of the world in dynamic environments where changes take place regularly.

Related to OFC function in decision making and reinforcement learning, Wilson et al. (2014) proposed that OFC is involved in inferring the current state of the environment. For example, medial OFC had been shown to represent probability distribution on possible states of the environment (Chan et al., 2016), the current task state (Schuck et al., 2016) and uncertainty or entropy associated with the state of the environment (Muller et al., 2019). In the context of regime-shift detection, regimes can be regarded as states of the environment and therefore a change in regime indicates a change in the state of the environment. Muller et al. (2019) found that in dynamic environments where changes in the state of the environment happen regularly, medial OFC represented the level of uncertainty in the current state of the environment. Our finding that vmPFC represented individual participants' probability estimates of regime shifts suggest that vmPFC and/or OFC are involved in inferring the current state of the environment through estimating whether the state has changed. Our finding that vmPFC represented individual participants' sensitivity to transition probability further suggest that vmPFC and/or OFC contribute to individual participants' biases in state inference (over- and underreactions to change) in how these brain areas respond to the volatility of the environment.

Our results are also closely related to the literature on the neural mechanisms for evidence accumulation in decision making (Gold & Shadlen, 2007; Mante et al., 2013; Philiastides et al., 2010; Roitman & Shadlen, 2002; Yates et al., 2017). In our task, evaluating the signals (red or blue balls) and in particular, the strength of change evidence associated with the signals is central to performing the task. Normatively, such evaluation should depend on the signal diagnosticity. In a highly diagnostic environment, seeing a red ball should signal a strong possibility of being in the red regime, while seeing a blue ball should signal otherwise. By contrast, in a low diagnostic environment, a red (resp. blue) ball is not strongly indicative of a red (resp. blue) regime. Hence, the evaluation of signals should reflect the interaction between the signals and the diagnosticity of the signals.

We found that this key computation was implemented in a frontoparietal network commonly referred to as the frontoparietal control network (Buckner et al., 2013; Dosenbach et al., 2007; Seeley et al., 2007; Vincent et al., 2008; Yeo et al., 2011). This network was proposed to support adaptive control functions, including initiating control and providing flexibility to adjust the level of control through feedback (Dosenbach et al., 2007). The IPS and dlPFC, part of this network, have also been found to play a major role in the top-down control of attention (Corbetta & Shulman, 2002; Woldorff et al., 2004). In perceptual decision making, the IPS and dlPFC were also shown to represent the accumulation of sensory evidence that leads to the formation of a perceptual decision (Heekeren et al., 2004; Heekeren et al., 2006). Our findings—that activity in this network does not reflect just the sensory signals (red or blue balls) but how these signals should be interpreted through the lens of their diagnosticity—highlights the involvement of the frontoparietal control network in computing the strength of evidence through combining information about signals and knowledge about the precision of those signals.

For the frontoparietal network, we identified its involvement in our task through finding that its activity correlated with strength of change evidence (Fig. 4) and individual subjects' sensitivity to signal diagnosticity (Fig. 5). Conceptually, these two findings reflect how individuals interpret the signals (signals consistent or inconsistent with change) in light of signal diagnosticity. This is because (1) strength of change evidence is defined as signals (+1 for signal consistent with change, and -1 for signal inconsistent with change) multiplied by signal diagnosticity and (2) sensitivity to signal diagnosticity reflects how individuals subjectively evaluate signal diagnosticity. At the theoretical level, these two findings can be interpreted through our computational framework in that both the strength of change evidence and sensitivity to signal diagnosticity contribute to estimating the likelihood of change (Equations 1 and 2 in *Methods*).

Our result on the intraparietal sulcus (IPS) being part of the brain network that represents diagnosticity-weighted sensory signals are consistent with previous studies showing that IPS is involved in accumulating sensory evidence over time (Gold & Shadlen, 2007 [↗](#)). There are three interesting aspects of our data that add to the current literature on evidence accumulation. First, IPS representations for sensory evidence need not be in the space of actions. Unlike previous studies showing that IPS represents sensory evidence for potential motor actions, we found that IPS represents the strength of evidence in favor of or against regime shifts. This result points to a more general role of the IPS in estimating the strength of sensory evidence. In fact, our result suggests that it depends on the task goal, which in the current study is to estimate whether a change has taken place. Second, although evidence accumulation is important and necessary for a wide array of cognitive functions, it is not a central requirement for the regime-shift task. Bayesian updating—the framework in which our system-neglect model was built upon—only requires the computation of strength of change evidence associated with the signal shown in the latest period. By showing that IPS representing this quantity, this suggests that IPS is involved in evaluating the latest piece of evidence necessary for belief updating.

In the current study, the central opercular cortex—in addition to the vmPFC—is another brain region that represented the probability estimates of change. Like the vmPFC, activity in this region negatively correlated with the probability estimates of change. This finding is associated with previous findings on change detection using the oddball paradigm. Using the oddball paradigm, it was found that the central opercular cortex is involved in the detection of change, showing stronger activation in blocks containing only the standard stimulus than blocks containing both the standard and deviant stimulus (Hedge et al., 2015 [↗](#)) and correlating with ERP P3 signals at the trial-level that reflected differences between standard and deviant stimuli (Warbrick et al., 2009 [↗](#)). There are two implications here. First, our findings suggest that the central opercular cortex is not only involved in the detection of change—as revealed by the oddball tasks—but also is involved in the estimation of change where there is uncertainty regarding whether the state of the world had changed. Second, the central opercular cortex may be part of a common pathway for the detection of change across very different tasks such as the oddball paradigm and our regime-shift detection task.

In the current study, our psychometric-neurometric analysis focused on comparing behavioral sensitivity with neural sensitivity to the system parameters (transition probability and signal diagnosticity). We measured sensitivity by estimating the slope of behavioral data (behavioral slope) and neural data (neural slope) in response to the system parameters. Previous studies had adopted a similar approach (Ting et al., 2015a [↗](#); Vilares et al., 2012 [↗](#); Yang & Wu, 2020 [↗](#)). For example, Vilares et al. (2012) [↗](#) found that sensitivity to prior information (uncertainty in prior distribution) in the orbitofrontal cortex (OFC) and putamen correlated with behavioral measure of sensitivity to the prior. In the current study, transition probability acts as prior in the system-neglect framework (Eq. 2 [↗](#) in *Methods*) and we found that ventromedial prefrontal cortex represents subjects' sensitivity to transition probability. Together, these results suggest that OFC (with vmPFC being part of OFC, see Wallis, 2011) is involved in the subjective evaluation of prior information in both static (Vilares et al., 2012 [↗](#)) and dynamic environments (current study). In addition, distinct from vmPFC in representing sensitivity to transition probability or prior, we found through the behavioral-neural slope comparison that the frontoparietal network represents how sensitive individual decision makers are to the diagnosticity of signals in revealing the true state (regime) of the environment.

In summary, our results suggest that an important mechanism for under- and overreactions to change has to do with neural sensitivity to system parameters that impact regime shifts. Importantly, different system parameters appear to recruit distinct brain networks according to their unique computational specializations. Given that under- and overreactions underly a wide array of human judgments, our findings indicate that network-level computational specificity and parameter selectivity are two key building blocks that give rise to human judgment biases.

Materials and methods

The data and analysis code are available at <https://osf.io/xh7dy/>.

We performed three fMRI experiments (90 subjects in total, 30 subjects for each experiment). Experiment 1 was the main experiment where we investigated the neurocomputational substrates for regime shifts. Experiments 2 and 3 were control experiments. Experiment 2 was designed to rule out brain activity that correlated with probability estimates but was not specifically about regime shifts. Experiment 3 attempted to rule out brain activity that correlated with entering numbers through button presses. In the main text, we only presented the results from Experiment 1. The procedure and results of Experiments 2 and 3 were presented in Supplementary Materials.

Subjects

All subjects gave informed written consent to participate in the study. All subjects were right-handed. The study procedures were approved by the National Yang Ming Chiao Tung University Institutional Review Board. Ninety subjects participated in this study:

- Experiment 1: $n = 30$ subjects; 15 males; mean age: 22.9 years; age range: 20 to 29 yrs.
- Experiment 2: $n = 30$ subjects; 15 males; mean age: 23.3 years; age range: 20 to 30 years.
- Experiment 3: $n = 30$ subjects; 15 males; mean age: 23.7 years; age range: 20 to 34 years.

Subjects were paid 300 New Taiwan dollar (TWD, 1 USD = 30 TWD) for participating the behavioral session and 500 TWD for the fMRI session. Subjects received additional monetary bonus based on his or her performance on probability estimation in Experiments 1 and 2 (Experiment 1: an average of 209 and 212 TWD for the behavioral and fMRI session respectively; Experiment 2: an average of 223 and 206 TWD for the behavioral and fMRI session respectively). In Experiment 3, subjects received the bonus based on their performance for entering the correct number (an average of 243 TWD for the fMRI session).

Procedure

Overview

Experiment 1 consisted of two sessions—a behavioral session followed by an fMRI session—that took place on two consecutive days. Subjects performed the same task in both sessions. The goals of having the behavioral session were to familiarize subjects with the task and to have enough trials—along with the fMRI session—to reliably estimate the parameters of the system-neglect model. Details of Experiments 2 and 3 can be found in the Supplement.

Regime-shift detection task

In this task, the environments the subjects are described as *regimes*. There were two possible regimes and at any point in time the regime can shift from one to another. Subjects judged whether the regime had shifted based on three sources of information: transition probability, signal diagnosticity, and signals. Prior to each trial, the subjects were given information about the transition probability and signal diagnosticity of the upcoming trial. The transition probability described how likely it is that the regime can shift from one to another. The signal diagnosticity described how different the two regimes are. The two regimes—red and blue—were represented by a red urn and a blue urn respectively, with each urn consisting of red and blue balls. The red urn had more red balls and the blue urn had more blue balls. The two urns always shared the same ratio of the dominant ball color to the dominated ball color. That is, the ratio of the red balls to the blue balls in the red urn was same as the ratio of the blue balls to the red balls in the blue urn. Signal diagnosticity was quantitatively defined by this ratio. For example, when the ratio was 9:1, signal diagnosticity (d) was 9. Under this definition, larger signal diagnosticity indicated that the two regimes were more different from each other.

After information about the transition probability and signal diagnosticity were revealed, 10 sensory signals—red or blue balls shown on the computer screen—were sequentially presented to the subjects. Each signal was sampled with replacement from one of the two urns. We also

referred to the sequential presentation of these signals as *periods*. By design, each trial always started at the red regime, but the regime can shift from the red to the blue regime at any point in time during a trial, including prior to the sampling of the first signal. In addition, also by design, the regime can shift only once in a trial (i.e., the blue regime was an absorbing or trapping state). That is, once the regime changed from the red to the blue regime, signals would be sampled from the blue urn until the end of the trial. Subjects provided an estimate of the probability that current regime was the blue regime at each period when a new signal was presented. Since the regime can only shift from the red to the blue, this probability estimate was equivalent to the probability estimate of regime shift.

Stimuli

For each subject and each trial separately, we generated the stimuli, i.e., the sequence of red and blue balls according to the transition probability and signal diagnosticity of the trial. Before the start of each period, we first determined whether the regime would shift from the Red regime (the starting regime) to the Blue regime by sampling from the transition probability. There were two possible outcomes: 1 indicates a change in regime, whereas 0 indicates no change. If the outcome were 1, we would sample from the Blue regime for that period and for all the remaining period(s). If the outcome were 0, we would sample from the Red regime for that period and repeat the same process described above for the next period.

Manipulations of transition probability and signal diagnosticity

We implemented a 3×3 factorial design where there were three possible transition probabilities, denoted as q where $q = [0.01, 0.5, 0.1]$ and three levels of signal diagnosticity, denoted as d where $d = [1.5, 3, 9]$. The transition probability—the probability that the regime shifts from the red to the blue regime—was used to simulate the stability of the environment. Larger transition probabilities indicate less stable environments. The signal diagnosticity was the ratio of dominant balls to the dominated balls in the urns. For example, when $d = 9$, it indicates that the red balls are 9 times more than the blue balls in the red urn and that the blue balls are 9 times more than the red balls in the blue urn. In this case, the two regimes were highly different and hence the signals shown to the subjects should be highly diagnostic of the regime the subjects were currently in.

Session 1: behavioral session

A pre-test block followed by the main task (8 blocks of trials) were implemented in the behavioral session. We implemented the pre-test block to make sure the subjects understood the experimental instructions and to train them to become familiar with the task, especially the time limit in entering probability estimates (see Trial sequence below for details on the pre-test block). A 3 (transition probability) by 3 (signal diagnosticity) factorial design was implemented, resulting in a total of 9 experimental conditions. Each block consisted of 9 trials (one for each condition, randomized in order) and each trial consisted of 10 periods or equivalently, 10 sequentially presented signals. After the pre-test block, subjects performed the main task that consisted of 8 blocks of trials. The behavioral session took approximately 70 minutes to complete.

Trial sequence

At the beginning of each trial, the subjects were shown information about the transition probability and signal diagnosticity (3s) (Fig. 1A [↗](#)). This was followed by the sequential presentation of 10 signals. At each period, a signal—a red or blue dot sampled from the current regime was shown. Subjects' task was to estimate the probability that the current regime was the blue regime within 4s. The subjects indicated the probability estimate through two button presses. During the experiment, the subjects placed his or her 10 fingers on a 10-button keypad. Each finger was associated with a single button that corresponded to a unique integer value from 0 to 9, with the left little finger for 1, left ring finger for 2, left middle finger for 3, left index finger for 4, left thumb for 5, right thumb for 6, right index finger for 7, right middle finger for 8, right ring finger for 9, and right little finger for 0. To enter the probability estimate, the subjects had to first press the number corresponding to the tens and second the number corresponding to the ones.

For example, if the probability estimate was 95%, the subjects first had to press the “9” button and second the “5” button. Once the subjects entered the probability estimate, she or he was not allowed to change it. After providing the probability estimate, they were given a brief feedback (0.5s) on the number they just entered. If they failed to indicate probability estimate within the time limit (4s), a “too slow” message would be shown. At the end of each trial, the subjects received feedback (2s) on the amount of monetary bonus earned in the trial (2s) and information about whether the regime shifted during the trial. If the regime was shifted, the signal that was drawn right after the shift took place would be highlighted in white. This was followed by a variable inter-trial interval (ITI, 1s to 5s in steps of 1s drawn from a discrete uniform distribution).

We implemented a pre-test block at the beginning of the behavioral so as to train the subjects to enter the probability estimate within the time limit (4s). In the pre-test block, we started with a lenient time limit and gradually decreased it. The time limit for the first three trials was 20s, 10s from trial number 4 to 6, and 4s from trial number 7 to 9. Recall that in each trial, the subjects encountered 10 sequentially presented signals (red and blue balls) and had to provide probability estimate at each period. The subjects therefore entered 30 probability estimates under each of these three time limits—from 20s, 10s, to 4s. After the pre-test block, all the subjects were able to enter the probability estimates within 4s.

Session 2: fMRI session

The fMRI session consisted of 3 blocks (9 trials in each block, with each trial consisting of 10 sequentially presented signals). The task was identical to the behavioral session (except with varying inter-stimulus interval for the purpose of fMRI analysis) and took approximately 60 minutes to complete. The subjects indicated the probability estimate through two button presses. During the experiment, the subjects placed his or her 10 fingers on a 10-button keypad. Each finger was associated with a single button that corresponded to a unique integer value from 0 to 9 (starting from the left pinkie for 1, left ring finger for 2, left middle finger for 3, and etc., to right ring finger for 9, and finally right pinkie for 0). The trial sequence was identical to the behavioral session with a few exceptions. First, each new signal was presented for 4s regardless of when subjects make a response. Second, we added a variable inter-stimulus interval (ISI, 1s to 5s in steps of 1 drawn from a uniform distribution) between two successive signals. We also made the range of ITI to be slightly larger (1s to 7s drawn from a discrete uniform distribution between in steps of 1s) than the behavioral session. The design of variable ISIs and ITIs was to allow better dissociations between events, i.e., between different sensory signals presented during a trial, and between trials for fMRI analysis.

Monetary bonus

To incentivize subjects to perform well in this experiment, they received monetary bonus based on his or her probability estimates. The bonus rule was designed so that the subjects who gave more accurate estimates would earn more bonus. The bonus structure used a quadratic payoff:

$$Bonus(t) = \$30 \times (0.1 - 0.2 \times (P_t - B_t)^2),$$

where P_t is the probability estimate that the current regime was blue at the t -th period in a trial and B_t is the regime at t ($B_t = 1$ for the blue regime and $B_t = 0$ for the red regime). For each probability estimate, the bonus therefore ranged from winning \$3 to losing \$3 TWD. For example, if the subject gave a 99% probability estimate that the current regime was blue and the current regime was indeed the blue regime, she would receive a bonus close to 3 TWD. By contrast, if the subjects gave a 1% probability estimate that the current regime was blue but the current regime was the red regime, she would receive a penalty close to 3 TWD. With 10 probability estimates given in each trial, the subjects can therefore receive a bonus up to 30 TWD or a penalty up to 30 TWD in a trial. The subjects did not receive feedback on bonus after each probability estimate. Instead, at the end of each trial, the subjects received information about the total amount won or lost in that trial. The final total bonus was realized by randomly selecting 10 trials at the end of the experiment.

Computational models for regime shift

We examined two computational models for regime shift: the Bayesian model and the system-neglect model (Massey & Wu, 2005). The Bayesian model was parameter-free and was used to compare with subjects' probability estimates of regime shift. The system-neglect model is a quasi-Bayesian model or parameterized version of the Bayesian model that was fit to the subjects' probability-estimate data. The parameter estimates of the system-neglect model were further used in the fMRI analysis so as to identify neural representations for over- and underreactions to change.

Bayesian model

Here we describe the Bayesian posterior odds of shifting to the blue regime given the period history H_t (Edwards, 1968; Massey & Wu, 2005):

$$\frac{P_t^B}{1-P_t^B} = \frac{\Pr(B_t|H_t)}{\Pr(R|H_t)} = \frac{1-(1-q)^t}{(1-q)^t} \sum_{j=1}^t \frac{q(1-q)^{j-1}}{1-(1-q)^t} d^{t+1-j-2\sum_{k=j}^t r_k}, \quad (1)$$

$P_t^B = \Pr(B_t|H_t)$ is the posterior probability that the regime has shifted to the blue regime at the t -th period, H_t denotes the sequence of history from r_1 to r_t . Here r_t denotes the t -th period, where $r_t = 1$ when the signal at t is red, and $r_t = 0$ when the signal at t is blue. The transition probability and signal diagnosticity are denoted by q and d respectively.

The posterior odds are the product of the prior odds and the likelihood ratio. The prior odds, $\frac{1-(1-q)^t}{(1-q)^t}$, indicates that given the transition probability q , the probability that the regime has shifted at time t , $1 - (1 - q)^t$, relative to the probability of no change in regime, $(1 - q)^t$. The likelihood ratio, $\sum_{j=1}^t \frac{q(1-q)^{j-1}}{1-(1-q)^t} d^{t+1-j-2\sum_{k=j}^t r_k}$, indicates the probability of observing the history of signals H_t given that the regime has shifted relative to the probability given that the regime has not shifted. This requires considering all the possibilities on the timing of shift, i.e., the likelihood ratio that the regime was shifted at $t = j$, $d^{t+1-j-2\sum_{k=j}^t r_k}$, weighted by its odds $\frac{q(1-q)^{j-1}}{1-(1-q)^t}$.

Since these possibilities are disjoint events, the likelihood ratio that the regime has shifted is simply the weighted sum of all the likelihood ratios associated with these disjoint possibilities.

Index of overreaction to change

To quantify under- and overreactions to change, we derived an Index of Overreaction (IO). In short, this index reflects the degree to which the subjects overreacted or underreacted to change such that an index value greater than 0 indicates overreaction and an index value smaller than 0 suggests underreaction. To compute IO , we compared the change in probability estimates between two adjacent periods ($\Delta P_t = P_t - P_{t-1}$) with the normative change in probability estimates according to the Bayesian model ($\Delta P_t^B = P_t^B - P_{t-1}^B$). Here we use P to denote the subject's probability estimate at the t -th period and P_t^B to denote the Bayesian probability estimate. We computed IO ($IO = \Delta P_t - \Delta P_t^B$) separately for each subject and for each condition or combination of transition probability and signal diagnosticity. Overreaction is defined when the actual change in probability estimates is greater than the normative change in probability estimates ($\Delta P_t > \Delta P_t^B$). By contrast, underreaction is defined when the actual change is smaller than the normative change in probability estimate ($\Delta P_t < \Delta P_t^B$). When $\Delta P_t^e = \Delta P_t^B$, there neither overreaction nor underreaction to change.

System-neglect model

Following Massey and Wu (2005), we fit the system-neglect model—a quasi-Bayesian model—to the subjects' probability estimates. In this model, we aimed to capture system-neglect—that people respond primarily to the signals and secondarily to the system generating the signals. Responding secondarily to the system indicates that people with system-neglect would be less sensitive to the

changes in the system parameters compared with the normative Bayesian. This is captured by adding a weighting parameter to each system parameter level. Hence, two weighting parameters, α and β , were added to Eq. (1) to transition probability and signal diagnosticity respectively such that

$$\frac{P_t}{1-P_t} = \frac{\Pr(B_t|H_t)}{\Pr(R|H_t)} = \frac{1-(1-\alpha q)^t}{(1-\alpha q)^t} \sum_{j=1}^t \frac{q(1-q)^{j-1}}{1-(1-q)^t} d^{\beta(t+1-j-2\sum_{k=j}^t r_k)} \quad (2)$$

where P_t is the probability estimate that regime has changed to the blue regime at period t . We separately estimated α for each level of transition probability and β for each level of signal diagnosticity. This was implemented by setting dummy variables for each level of transition probability and signal diagnosticity in Eq. (2)

$$\alpha = \alpha_1 Q_1 + \alpha_2 Q_2 + \alpha_3 Q_3 \quad \text{and} \quad \beta = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3.$$

Where Q_i is the dummy variable for transition probability q_i and D_j is the dummy variable for diagnosticity d_j . α_i and β_j therefore respectively reflect the sensitivity to different levels of transition probability and signal diagnosticity. For each subject separately, we performed nonlinear regression (using the *fitnlm* function in MATLAB) to fit the model to the subject's probability estimates and estimated the parameters of interest.

Parameter recovery analysis

To examine whether the fitting procedure gave reasonable parameter estimates of the system-neglect model, we performed a parameter recovery analysis (Wilson & Collins, 2019). The analysis proceeded in the following steps. First, we simulated each subject's probability estimation data based on the system-neglect model by using the parameter estimates obtained for that subject after fitting the system-neglect model to the subject's probability estimation data. Second, we fit the system-neglect model to the simulated data. Third, as a measure of parameter recovery, we computed the correlation across subjects between the estimated parameters and the parameter values we used to simulate data, where larger correlation indicates better recovery.

Fourth, we repeated the above steps by adding independent white noise to the simulated data, from a Gaussian distribution with mean 0 and variance σ^2 . We implemented 5 levels of noise with $\sigma_{noise} = \{0.01, 0.05, 0.1, 0.2, 0.3\}$ and examined the impact of noise on parameter recovery. These noise levels covered the range of empirical noise levels estimated from the subjects. To estimate each subject's noise level, we incorporated a noise parameter into the system-neglect model. We assumed that probability estimates are noisy and modeled them with a Gaussian distribution where the noise parameter (σ_{noise}) is the standard deviation. At each period, a probability estimate of regime shift was computed according to the system-neglect model where Θ is the set of parameters including parameters in the system-neglect model and the noise parameter. The likelihood function, $L(\Theta)$, is the probability of observing the subject's probability estimate at period t , p_t , given Θ , $L(\Theta) = P(p_t|\Theta)$. Since we modeled the noisy probability estimates with a Gaussian distribution, we can therefore express $L(\Theta)$ as $L(\Theta) \sim N(p_t; p_t^{SN}, \sigma_{noise})$ where p_t^{SN} is the probability estimate predicted by the system-neglect (SN) model at period t . As a reminder, we referred to a 'period' as the time when a new signal appeared during a trial (for a given transition probability and signal diagnosticity). To find the maximum likelihood estimates of the parameters, Θ_{MLE} , we summed over all periods the negative natural logarithm of likelihood and used MATLAB's *fmincon* function to find Θ_{MLE} . Across subjects, we found that the mean noise estimate was 0.1735 and ranged from 0.1118 to 0.2704 (Fig. S3 in SI).

Impact of noise homoscedasticity on parameter estimation

The *fitnlm* function we used in MATLAB to estimate the parameters of the system-neglect models assumes homoscedasticity—that noise is constant when fitting the model. However, it is possible that this assumption was violated, that subjects' actual noise was heteroscedastic, for example, having larger noise when probability estimates are around 0.5 and smaller noise at the two extremes (0 and 1). To examine this possibility, for each subject, we divided probability into 5

intervals ([0.0–0.2), [0.2–0.4), [0.4–0.6), [0.6–0.8), and [0.8–1.0]) and computed the residual standard deviation separately for each bin. Here the residual was the difference between subjects' probability estimates (p_t) and the probability estimates derived from the system-neglect model based on each subject's parameter estimates (p_t^{SN}) $residual = p_t - p_t^{SN}$. We found that homoscedasticity was indeed violated—the standard deviation of residuals was smallest when probability was 0.1 and increased as a function of probability from 0.1 to 0.5. When probability was larger than 0.5, the residual standard deviation was similar (Fig. S4A in *SI*). To see how this would affect parameter estimation, we performed parameter recovery analysis assuming heteroscedasticity. That is, we simulated subjects' probability estimates using the empirically estimated, probability-dependent residual standard deviation as the standard deviation of the Gaussian noise and using the simulated data to estimate parameters in the system-neglect model. We found that we were able to recover the parameters well (Fig. S4B in *SI*) and the result was similar to the parameter recovery assuming homoscedastic noise (Fig. S3 in *SI*). This suggested that heteroscedastic noise did not impact the accuracy of parameter estimation when homoscedasticity was assumed.

fMRI data acquisition

All the imaging parameters, including the EPI sequence and MPAGE sequence, remained the same throughout the three experiments. For Experiments 1 and 2, the subjects completed the task in a 3T Siemens MRI scanner (MAGNETOM Trio) equipped with a 32-channel head array coil. Experiment 3 was collected at a later date after the same scanner went through a major upgrade (from Trio to the Prisma system) where a 64-channel head array coil was used. Each subject completed three functional runs. Before each run, a localizer scan was implemented for slice positioning. For each run, T2*-weighted functional images were collected using an EPI sequence (TR=2000ms, TE=30ms, 33 oblique slices acquired in ascending interleaved order, 3.4×3.4×3.4 mm isotropic voxel, 64×64 matrix in 220 mm field of view, flip angle 90°). To reduce signal loss in the ventromedial prefrontal cortex and orbitofrontal cortex, the sagittal axis was tilted clockwise up to 30°. Each run consisted of 9 trials and a total of 374 images. After the functional scans, T1-weighted structural images were collected (MPAGE sequence with TR=2530ms, TE=3.03ms, flip angle=7°, 192 sagittal slices, 1×1×1 mm isotropic voxel, 224×256 matrix in a 256-mm field of view). For each subject, a field map image was also acquired for the purpose of estimating and partially compensating for geometric distortion of the EPI image so as to improve registration performance with the T1-weighted images.

fMRI preprocessing

The imaging data were preprocessed with FMRIB's software Library (FSL version 6.0). First, for motion correction, MCFLIRT was used to remove the effect of head motion during each run. Second, FUGUE (FMRIB's Utility for Geometric Unwarping of EPIs) was used to estimate and partially compensate for geometric distortion of the EPI images using field map images collected for the subject. Third, spatial smoothing was applied with a Gaussian kernel with FWHM=6mm. Fourth, a high-pass temporal filtering was applied using Gaussian-weighted least square straight-line fitting with $\sigma = 50$ s. Fifth, registration was performed in a two-step procedure, with the field map used to improve the performance of registration. First, EPI images were registered to the high-resolution brain T1-weighted structural image (non-brain structures were removed via FSL's BET (Brain Extraction Tool)). Second, the transformation matrix (12-parameter affine transformation) from the T1-weighted image to the Montreal Neurological Institute (MNI) template brain was estimated using FLIRT (FMRIB's Linear Image Registration Tool), followed by nonlinear registration using FNIRT (FMRIB's Non-linear Image Registration Tool) with a 10mm warp resolution. This two-step procedure allowed for transforming the EPI images to the standard MNI template brain.

General Linear Models of BOLD signals

All GLM analyses were carried out in the following steps (Beckmann et al., 2003). First, BOLD time series were pre-whitened with local autocorrelation correction. A first-level FEAT analysis was carried out for each run of each subject. Second, a second-level (subject-level) fixed-effect (FE) analysis was carried out for each subject that combined the first-level FEAT results from different runs using the summary statistics approach. Finally, a third-level (group-level) mixed-effect (ME) analysis using FSL's FLAME module (FMRIB's Local Analysis of Mixed Effects) was carried out across subjects by taking the FE results from the previous level and treating subjects as a random effect (Woolrich et al., 2004). All reported whole-brain results were corrected for multiple comparisons. We first identified clusters of activation by defining a cluster-forming threshold of the z statistic ($z > 3.1$ or equivalently $p < .001$) (Eklund et al., 2016; Woo et al., 2014). Then, a family-wise error corrected p -value of each cluster based on its size was estimated using Gaussian random field theory (Worsley et al., 1992). In addition, we performed nonparametric permutation test using the randomize function in FSL (threshold-free cluster enhancement or TFCE option; (Smith & Nichols, 2009) on all the contrasts reported.

GLM-1

This model was used for Fig 3. The model served two purposes. First, we used it to examine neural representations for probability estimates (P_t) and belief revision (ΔP_t) (Fig. 3B). Second, we used this model to compare results between the main experiment (Experiment 1) and the control experiments (Experiments 2 and 3) (Fig. 3CD). We implemented the following regressors. At the time of each signal presentation, we implemented the following regressors: (R1) an indicator regressor with length equal to the subject's RT, (R2) R1 multiplied by the subject's probability estimate, P_t , that the signal came from the Blue regime, (R3) R1 multiplied by difference in the subject's probability estimate between two successive periods, ΔP_t , which captured the updating of belief about change, (R4) R1 multiplied by the degree of certainty in probability estimate $|P_t - 0.5|$, (R5) R1 multiplied by the period number (from 1 to 10). Both positive and negative contrasts of R2 (P_t) and R3 (ΔP_t) were set up to identify activity that either positively or negatively correlate with these regressors. At the end of each trial (after subjects saw all 10 signals), we provided the monetary bonus the subject earned in that trial. We implemented an indicator regressor (R6) and a parametric regressor for the subject's winning (R7). We implemented an indicator regressor (length equal to 4s) for the no-response periods (R8), which corresponded to the period(s) where the subject did not indicate the probability estimate within the time limit (4s). Finally, to directly address the motor confound issue, we implemented an action-handedness regressor, (R9), which was R1 multiplied by action-handedness. This regressor served to address the motor confounds of probability estimates P_t , as higher probability estimates preferentially involved right-handed responses for entering higher digits and lower estimates involved left-handed responses. Therefore, at the time of each signal presentation, the action-handedness regressor coded -1 if both finger presses to enter P_t involved left hand, 0 if using one left finger and one right finger, and 1 if both finger presses involved right hand. For Experiment 1, we also performed a GLM analysis that was identical to GLM-1 except that it did not include the action-handedness regressor and the results on both P_t and ΔP_t were largely identical (Supplementary Table S10). Note that for both Experiment 1 (the main regime-shift experiment) and Experiment 2 (control experiment), P_t indicates the probability estimate that the signal came from the Blue regime. However, for Experiment 3 (control experiment) subjects were instructed to press a two-digit number shown on the screen at each period. Hence, the P_t regressor for Experiment 3 was effectively the instructed number, and the ΔP_t regressor was effectively the difference in instructed number between adjacent periods.

GLM-2

This model was implemented to examine the effects of component variables critically contributing to regime-shift probability estimation. In particular, we used this model to identify the effects on the strength of change evidence (based on R11 below) and the intertemporal prior (based on R9 below) (Fig. 4 in main text). The model was identical to GLM-1 from (R1) to (R8) with the

addition of the following regressors: (R9) R1 multiplied by the intertemporal prior, defined as $\ln\left(\frac{1-(1-q)^t}{(1-q)^t}\right)$, where q is the transition probability and t is the period, (R10) R1 multiplied by the natural logarithm of signal diagnosticity, $\ln(d)$, (R11) R1 multiplied by the current signal (1 for blue ball, -1 for red ball), and (R12) the interaction $R9 \times R10$ (interaction between signal diagnosticity and signal). We note that since these component variables (R9 to R12) contributed to regime-shift probability estimation, they were correlated with P_t and ΔP_t . Having both with P_t and ΔP_t in the same model as these component variables would therefore introduce collinearity and reduce the reliability of regression coefficients. While we were aware of this issue, we also recognized that having these correlated regressors in the same model have the advantage of statistically stating that neural activity attributed to a particular variable cannot be otherwise attributed to another variable that correlated with it. We also performed another GLM analysis that was identical to GLM-2 except that it did not include P_t and ΔP_t in the model and the results on strength of change evidence and intertemporal prior shown in Fig. 4 were largely identical (Supplementary Table S11).

GLM-3

This model was the basis of results shown in Fig. 5. We set up two sets of regressors, one set for when the blue signal (signal for potential change) appeared and the other for when the red signal (signal for no change) appeared. For each set, at the time of signal presentation, we included 9 indicator regressors, one for every combination of 3 transition probability levels and 3 signal diagnosticity levels. Based on these 9 regressors, we set up a linear-increasing contrast separately for signal diagnosticity and transition probability. At the subject level (second level), the parameter estimate of these contrasts reflects individual subjects' sensitivity to signal diagnosticity and transition probability were extracted from a given ROI and was used to correlate with individual subjects' behavioral sensitivity to signal diagnosticity and transition probability. The ROIs used were vmPFC, ventral striatum, dmPFC, bilateral IFG, and bilateral IPS. These ROIs were identified by results from GLM 1 and 2 and were constructed in a statistically independent manner using leave-one-subject-out method. See Independent region-of-interest (ROI) analysis in the section below. Same as in the previous two GLMs, at the end of each trial when feedback on the current-trial monetary bonus was revealed, we implemented an indicator regressor and a parametric regressor for the monetary bonus. Finally, we implemented an indicator regressor (length equal to 4s) for no-response periods. These corresponded to the period(s) where the subject did not indicate the probability estimate within the time limit (4s).

Independent regions-of-interest (ROIs) analysis

We performed two kinds of independent ROI analysis—leave-one-subject-out (LOSO) method and functional/structural masks based on previous meta-analysis paper (Bartra et al. (2013) or existing structural atlases. Specifically, we used the functional mask in vmPFC from Bartra et al. (2013) and structural mask for the ventral striatum based on the Harvard-Oxford cortical and subcortical atlases in FSL. For LOSO, based on results from the whole-brain analysis, we created independent and unbiased ROIs using the leave-one-subject-out (LOSO) method (Litt et al., 2011; Ting et al., 2015b). The LOSO method was used to analyze probability estimates and representations for transition probability and signal diagnosticity described above. We performed the analysis for each subject separately in the following steps. First, we identified the significant cluster in a brain region (e.g., dmPFC) that correlated with a contrast of interest (e.g., probability estimates) using all other subjects' data. We referred to this as the LOSO ROI. Second, we extracted the mean beta value (regression coefficient) within the LOSO ROI from the subject and used it for further statistical analysis. Note that the LOSO ROI tended to be different spatially between subjects. To give an idea of these differences and the spatial distribution of these ROIs, in Figs. 4 and 5 where the LOSO analysis was performed, we showed both the voxels that were part of the LOSO ROI of all the subjects in one color and the voxels that were part of at least one LOSO ROI in another color. Finally, we note that since the LOSO procedure involves, for each subject separately, performing the leave-one-subject-out inference at the group level and identifying the significant clusters of activation. Therefore, it is possible that some subject(s), after the LOSO inference, did

not have significant cluster in some brain region. For the analysis shown in Fig. 5C, for the left IFG ROI, we had 3 subjects that did not have a significant cluster in this brain region. All other ROIs had data from all subjects.

Parametric and nonparametric tests for difference in correlation coefficients

We implemented both parametric and nonparametric tests to examine whether the difference in Pearson correlation coefficients was significant. In the parametric test, we used the Fisher z transformation to transform the difference in correlation coefficients to the z statistic. That is, for two correlation coefficients, r_1 (with sample size n_1) and r_2 , (with sample size n_2), the z statistic of the difference in correlation is given by

$$z = \frac{r_1 - r_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

We referred to the correlation between neural and behavioral sensitivity at change-consistent (blue balls) signals as r_{blue} , and that at change-inconsistent (red balls) signals as r_{red} . For the Fisher z transformation, $r_1 = r_{blue}$ and $r_2 = r_{red}$. In the nonparametric test, we performed nonparametric bootstrapping to test for the difference in correlation (Efron & Tibshirani, 1994). That is, we resampled with replacement the dataset (subject-wise) and used the resampled dataset to compute the difference in correlation. We then repeated the above for 100,000 times so as to estimate the distribution of the difference in correlation coefficients, tested for significance and estimated p-value based on this distribution.

Acknowledgements

This work was supported by the National Science and Technology Council (NSTC) in Taiwan (Grants 108-2410-H-010-012-MY3, 110-2410-H-A49A-504-MY3 to S.-W.W.) and by the Brain Research Center, National Yang Ming Chiao Tung University from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. We acknowledge magnetic resonance imaging support from National Yang Ming Chiao Tung University, Taiwan, which is in part supported by the Ministry of Education plan for the top University.

Additional files

[Supplemental file](#)

Note

This reviewed preprint has been updated to correct a corresponding email address.

References

- Baker M., Wurgler J. (2007) Investor sentiment in the stock market. *Journal of economic perspectives* **21**:129-151
- Barberis N., Shleifer A., Vishny R. (1998) A model of investor sentiment. *Journal of financial economics* **49**:307-343
- Bartolo R., Averbeck B. B. (2020) Prefrontal Cortex Predicts State Switches during Reversal Learning. *Neuron* **106**:1044-1054 e1044 <https://doi.org/10.1016/j.neuron.2020.03.024>
- Bartra O., McGuire J. T., Kable J. W. (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**:412-427
- Beckmann C. F., Jenkinson M., Smith S. M. (2003) General multilevel linear modeling for group analysis in FMRI. *Neuroimage* **20**:1052-1063 <https://www.sciencedirect.com/science/article/pii/S105381190300435X>

- Behrens T. E., Woolrich M. W., Walton M. E., Rushworth M. F. (2007) Learning the value of information in an uncertain world. *Nat Neurosci* **10**:1214-1221 <https://doi.org/10.1038/nn1954>
- Benjamin D. J. (2019) Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1* **2**:69-186
- Brown S. D., Steyvers M. (2009) Detecting and predicting changes. *Cogn Psychol* **58**:49-67 <https://doi.org/10.1016/j.cogpsych.2008.09.002>
- Buckner R. L., Krienen F. M., Yeo B. T. (2013) Opportunities and limitations of intrinsic functional connectivity MRI. *Nat Neurosci* **16**:832-837 <https://doi.org/10.1038/nn.3423>
- Chan S. C., Niv Y., Norman K. A. (2016) A probability distribution over latent causes, in the orbitofrontal cortex. *Journal of Neuroscience* **36**:7817-7828
- Corbetta M., Shulman G. L. (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* **3**:201-215 <https://doi.org/10.1038/nrn755>
- Costa V. D., Tran V. L., Turchi J., Averbeck B. B. (2015) Reversal learning and dopamine: a bayesian perspective. *J Neurosci* **35**:2407-2416 <https://doi.org/10.1523/JNEUROSCI.1989-14.2015>
- Daniel K., Hirshleifer D., Subrahmanyam A. (1998) Investor psychology and security market under- and overreactions. *the Journal of Finance* **53**:1839-1885
- De Bondt W. F., Thaler R. (1985) Does the stock market overreact?. *the Journal of Finance* **40**:793-805
- Dosenbach N. U., Fair D. A., Miezin F. M., Cohen A. L., Wenger K. K., Dosenbach R. A., Fox M. D., Snyder A. Z., Vincent J. L., Raichle M. E., *et al.* (2007) Distinct brain networks for adaptive and stable task control in humans. *Proc Natl Acad Sci U S A* **104**:11073-11078 <https://doi.org/10.1073/pnas.0704320104>
- Edwards W. (1968) Conservatism in human information processing. In: *Formal representation of human judgment*
- Efron B., Tibshirani R. J. (1994) *An introduction to the bootstrap* Chapman and Hall/CRC.
- Eklund A., Nichols T. E., Knutsson H. (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences* **113**:7900-7905
- Fellows L. K., Farah M. J. (2003) Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain* **126**:1830-1837
- Gold J. I., Shadlen M. N. (2007) The neural basis of decision making. *Annu. Rev. Neurosci* **30**:535-574
- Griffin D., Tversky A. (1992) The weighing of evidence and the determinants of confidence. *Cognitive psychology* **24**:411-435
- Hedge C., Stothart G., Todd Jones J., Rojas Frias P., Magee K. L., Brooks J. C. (2015) A frontal attention mechanism in the visual mismatch negativity. *Behav Brain Res* **293**:173-181 <https://doi.org/10.1016/j.bbr.2015.07.022>
- Heekeren H. R., Marrett S., Bandettini P. A., Ungerleider L. G. (2004) A general mechanism for perceptual decision-making in the human brain. *nature* **431**:859-862 <https://doi.org/10.1038/nature02966>
- Heekeren H. R., Marrett S., Ruff D. A., Bandettini P. A., Ungerleider L. G. (2006) Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. *Proc Natl Acad Sci U S A* **103**:10023-10028 <https://doi.org/10.1073/pnas.0603949103>
- Izquierdo A., Brigman J. L., Radke A. K., Rudebeck P. H., Holmes A. (2017) The neural basis of reversal learning: An updated perspective. *Neuroscience* **345**:12-26 <https://doi.org/10.1016/j.neuroscience.2016.03.021>
- Kao C.-H., Khambhati A. N., Bassett D. S., Nassar M. R., McGuire J. T., Gold J. I., Kable J. W. (2020) Functional brain network reconfiguration during learning in a dynamic environment. *Nature communications* **11**:1682
- Kraemer C., Weber M. (2004) How do people take into account weight, strength and quality of segregated vs. aggregated data? Experimental evidence. *Journal of Risk and Uncertainty* **29**:113-142

- Kremer M., Moritz B., Siemsen E. (2011) Demand Forecasting Behavior: System Neglect and Change Detection. *Management Science* **57**:1827-1843 <https://doi.org/10.1287/mnsc.1110.1382>
- Litt A., Plassmann H., Shiv B., Rangel A. (2011) Dissociating Valuation and Saliency Signals during Decision-Making. *Cerebral Cortex* **21**:95-102 <http://cercor.oxfordjournals.org/content/21/1/95.abstract>
- Mante V., Sussillo D., Shenoy K. V., Newsome W. T. (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature* **503**:78-84
- Massey C., Wu G. (2005) Detecting regime shifts: The causes of under- and overreaction. *Management Science* **51**:932-947
- McGuire J. T., Nassar M. R., Gold J. I., Kable J. W. (2014) Functionally dissociable influences on learning rate in a dynamic environment. *Neuron* **84**:870-881
- Muller T. H., Mars R. B., Behrens T. E., O'Reilly J. X. (2019) Control of entropy in neural models of environmental state. *eLife* **8**:e39404 <https://doi.org/10.7554/eLife.39404>
- Nassar M. R., Wilson R. C., Heasly B., Gold J. I. (2010) An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience* **30**:12366-12378
- Nelson M. W., Bloomfield R., Hales J. W., Libby R. (2001) The effect of information strength and weight on behavior in financial markets. *Organizational Behavior and Human Decision Processes* **86**:168-196
- O'Doherty J., Kringelbach M. L., Rolls E. T., Hornak J., Andrews C. (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature neuroscience* **4**:95-102
- Payzan-LeNestour E., Bossaerts P. (2011) Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS computational biology* **7**:e1001048
- Payzan-LeNestour E., Dunne S., Bossaerts P., O'Doherty J. P. (2013) The neural representation of unexpected uncertainty during value-based decision making. *Neuron* **79**:191-201
- Philiastides M. G., Biele G., Heekeren H. R. (2010) A mechanistic account of value computation in the human brain. *Proceedings of the National Academy of Sciences* **107**:9430-9435
- Roitman J. D., Shadlen M. N. (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience* **22**:9475-9489
- Sanders N. R., Manrodt K. B. (2003) Forecasting Software in Practice: Use, Satisfaction, and Performance. *Interfaces* **33**:90-93 <https://doi.org/10.1287/inte.33.5.90.19251>
- Schoenbaum G., Chiba A. A., Gallagher M. (2000) Changes in functional connectivity in orbitofrontal cortex and basolateral amygdala during learning and reversal training. *Journal of Neuroscience* **20**:5179-5189
- Schonberg T., Fox C. R., Mumford J. A., Congdon E., Trepel C., Poldrack R. A. (2012) Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: an fMRI investigation of the balloon analog risk task. *Frontiers in neuroscience* **6**:25510
- Schuck N. W., Cai M. B., Wilson R. C., Niv Y. (2016) Human orbitofrontal cortex represents a cognitive map of state space. *Neuron* **91**:1402-1412
- Seeley W. W., Menon V., Schatzberg A. F., Keller J., Glover G. H., Kenna H., Reiss A. L., Greicius M. D. (2007) Dissociable intrinsic connectivity networks for salience processing and executive control. *J Neurosci* **27**:2349-2356 <https://doi.org/10.1523/JNEUROSCI.5587-06.2007>
- Seifert M., Ulu C., Guha S. (2023) Decision Making Under Impending Regime Shifts. *Management Science* <https://doi.org/10.1287/mnsc.2022.4661>
- Smith S. M., Nichols T. E. (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* **44**:83-98
- Soltani A., Izquierdo A. (2019) Adaptive learning under expected and unexpected uncertainty. *Nat Rev Neurosci* **20**:635-644 <https://doi.org/10.1038/s41583-019-0180-y>

- Ting C.-C.**, Yu C.-C., Maloney L. T., Wu S.-W. (2015a) Neural mechanisms for integrating prior knowledge and likelihood in value-based probabilistic inference. *Journal of Neuroscience* **35**:1792-1805
- Ting C.-C.**, Yu C.-C., Maloney L. T., Wu S.-W. (2015b) Neural Mechanisms for Integrating Prior Knowledge and Likelihood in Value-Based Probabilistic Inference. *The Journal of Neuroscience* **35**:1792-1805
- Tversky A.**, Griffin D., Heath C., Slovic P. (1990) Decision under conflict: Resolution and confidence in judgment and choice. Department of Psychology, Stanford University.
- Vilares I.**, Howard J. D., Fernandes H. L., Gottfried J. A., Kording K. P. (2012) Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology* **22**:1641-1648
- Vincent J. L.**, Kahn I., Snyder A. Z., Raichle M. E., Buckner R. L. (2008) Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. *J Neurophysiol* **100**:3328-3342 <https://doi.org/10.1152/jn.90355.2008>
- Walton M. E.**, Behrens T. E., Buckley M. J., Rudebeck P. H., Rushworth M. F. (2010) Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron* **65**:927-939
- Warbrick T.**, Mobascher A., Brinkmeyer J., Musso F., Richter N., Stoecker T., Fink G. R., Shah N. J., Winterer G. (2009) Single-trial P3 amplitude and latency informed event-related fMRI models yield different BOLD response patterns to a target detection task. *Neuroimage* **47**:1532-1544 <https://doi.org/10.1016/j.neuroimage.2009.05.082>
- Wilson R. C.**, Collins A. G. (2019) Ten simple rules for the computational modeling of behavioral data. *eLife* **8**:e49547 <https://doi.org/10.7554/eLife.49547>
- Wilson R. C.**, Takahashi Y. K., Schoenbaum G., Niv Y. (2014) Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**:267-279
- Woldorff M. G.**, Hazlett C. J., Fichtenholtz H. M., Weissman D. H., Dale A. M., Song A. W. (2004) Functional parcellation of attentional control regions of the brain. *J Cogn Neurosci* **16**:149-165 <https://doi.org/10.1162/089892904322755638>
- Woo C.-W.**, Krishnan A., Wager T. D. (2014) Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage* **91**:412-419
- Woolrich M. W.**, Behrens T. E. J., Beckmann C. F., Jenkinson M., Smith S. M. (2004) Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage* **21**:1732-1747 <http://www.sciencedirect.com/science/article/pii/S1053811903007894>
- Worsley K. J.**, Evans A. C., Marrett S., Neelin P. (1992) A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism* **12**:900-918
- Yang Y.-Y.**, Wu S.-W. (2020) Base rate neglect and neural computations for subjective weight in decision under uncertainty. *Proceedings of the National Academy of Sciences* **117**:16908-16919
- Yates J. L.**, Park I. M., Katz L. N., Pillow J. W., Huk A. C. (2017) Functional dissection of signal and noise in MT and LIP during decision-making. *Nat Neurosci* **20**:1285-1292 <https://doi.org/10.1038/nn.4611>
- Yeo B. T.**, Krienen F. M., Sepulcre J., Sabuncu M. R., Lashkari D., Hollinshead M., Roffman J. L., Smoller J. W., Zolke L., Polimeni J. R., et al. (2011) The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol* **106**:1125-1165 <https://doi.org/10.1152/jn.00338.2011>

Peer reviews

Reviewer #1 (Public review):

Summary:

The study examines human biases in a regime-change task, in which participants have to report the probability of a regime change in the face of noisy data. The behavioral results

indicate that humans display systematic biases, in particular, overreaction in stable but noisy environments and underreaction in volatile settings with more certain signals. fMRI results suggest that a frontoparietal brain network is selectively involved in representing subjective sensitivity to noise, while the vmPFC selectively represents sensitivity to the rate of change.

Strengths:

- The study relies on a task that measures regime-change detection primarily based on descriptive information about the noisiness and rate of change. This distinguishes the study from prior work using reversal-learning or change-point tasks in which participants are required to learn these parameters from experiences. The authors discuss these differences comprehensively.
- The study uses a simple Bayes-optimal model combined with model fitting, which seems to describe the data well. The model is comprehensively validated.
- The authors apply model-based fMRI analyses that provide a close link to behavioral results, offering an elegant way to examine individual biases.

Weaknesses:

The authors have adequately addressed most of my prior concerns.

My only remaining comment concerns the z-test of the correlations. I agree with the non-parametric test based on bootstrapping at the subject level, providing evidence for significant differences in correlations within the left IFG and IPS.

However, the parametric test seems inadequate to me. The equation presented is described as the Fisher z-test, but the numerator uses the raw correlation coefficients (r) rather than the Fisher-transformed values (z). To my understanding, the subtraction should involve the Fisher z-scores, not the raw correlations.

More importantly, the Fisher z-test in its standard form assumes that the correlations come from independent samples, as reflected in the denominator (which uses the n of each independent sample). However, in my opinion, the two correlations are not independent but computed within-subject. In such cases, parametric tests should take into account the dependency. I believe one appropriate method for the current case (correlated correlation coefficients sharing a variable [behavioral slope]) is explained here:

Meng, X.-l., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1), 172-175. <https://doi.org/10.1037/0033-2909.111.1.172> 

It should be implemented here:

Diedenhofen B, Musch J (2015) cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE* 10(4): e0121945. <https://doi.org/10.1371/journal.pone.0121945> 

My recommendation is to verify whether my assumptions hold, and if so, perform a test that takes correlated correlations into account. Or, to focus exclusively on the non-parametric test.

In any case, I recommend a short discussion of these findings and how the authors interpret that some of the differences in correlations are not significant.

<https://doi.org/10.7554/eLife.104684.2.sa2>

Reviewer #3 (Public review):

This study concerns how observers (human participants) detect changes in the statistics of their environment, termed regime shifts. To make this concrete, a series of 10 balls are drawn from an urn that contains mainly red or mainly blue balls. If there is a regime shift, the urn is changed over (from mainly red to mainly blue) at some point in the 10 trials. Participants report their belief that there has been a regime shift as a % probability. Their judgement should (mathematically) depend on the prior probability of a regime shift (which is set at one of three levels) and the strength of evidence (also one of three levels, operationalized as the proportion of red balls in the mostly-blue urn and vice versa). Participants are directly instructed of the prior probability of regime shift and proportion of red balls, which are presented on-screen as numerical probabilities. The task therefore differs from most previous work on this question in that probabilities are instructed rather than learned by observation, and beliefs are reported as numerical probabilities rather than being inferred from participants' choice behaviour (as in many bandit tasks, such as Behrens 2007 Nature Neurosci).

The key behavioural finding is that participants over-estimate the prior probability of regime change when it is low, and under estimate it when it is high; and participants over-estimate the strength of evidence when it is low and under-estimate it when it is high. In other words participants make much less distinction between the different generative environments than an optimal observer would. This is termed 'system neglect'. A neuroeconomic-style mathematical model is presented and fit to data.

Functional MRI results show that strength of evidence for a regime shift (roughly, the surprise associated with a blue ball from an apparently red urn) is associated with activity in the frontal-parietal orienting network. Meanwhile, at time-points where the probability of a regime shift is high, there is activity in another network including vmPFC. Both networks show individual differences effects, such that people who were more sensitive to strength of evidence and prior probability show more activity in the frontal-parietal and vmPFC-linked networks respectively.

Strengths

- (1) The study provides a different task for looking at change-detection and how this depends on estimates of environmental volatility and sensory evidence strength, in which participants are directly and precisely informed of the environmental volatility and sensory evidence strength rather than inferring them through observation as in most previous studies
- (2) Participants directly provide belief estimates as probabilities rather than experimenters inferring them from choice behaviour as in most previous studies
- (3) The results are consistent with well-established findings that surprising sensory events activate the frontal-parietal orienting network whilst updating of beliefs about the world ('regime shift') activates vmPFC.

Weaknesses

- (1) The use of numerical probabilities (both to describe the environments to participants, and for participants to report their beliefs) may be problematic because people are notoriously bad at interpreting probabilities presented in this way, and show poor ability to reason with this information (see Kahneman's classic work on probabilistic reasoning, and how it can be improved by using natural frequencies). Therefore the fact that, in the present study, people do not fully use this information, or use it inaccurately, may reflect the mode of information delivery.

(2) Although a very precise model of 'system neglect' is presented, many other models could fit the data.

For example, you would get similar effects due to attraction of parameter estimates towards a global mean - essentially application of a hyper-prior in which the parameters applied by each participant in each block are attracted towards the experiment-wise mean values of these parameters. For example, the prior probability of regime shift ground-truth values [0.01, 0.05, 0.10] are mapped to subjective values of [0.037, 0.052, 0.069]; this would occur if observers apply a hyper-prior that the probability of regime shift is about 0.05 (the average value over all blocks). This 'attraction to the mean' is a well-established phenomenon and cannot be ruled out with the current data (I suppose you could rule it out by comparing to another dataset in which the mean ground-truth value was different).

More generally, any model in which participants don't fully use the numerical information they were given would produce apparent 'system neglect'. Four qualitatively different example reasons are: 1. Some individual participants completely ignored the probability values given. 2. Participants did not ignore the probability values given, but combined them with a hyperprior as above. 3. Participants had a reporting bias where their reported beliefs that a regime-change had occurred tend to be shifted towards 50% (rather than reporting 'confident' values such 5% or 95%). 4. Participants underweighted probability outliers resulting in underweighting of evidence in the 'high signal diagnosticity' environment (10.1016/j.neuron.2014.01.020)

In summary I agree that any model that fits the data would have to capture the idea that participants don't differentiate between the different environments as much as they should, but I think there are a number of qualitatively different reasons why they might do this - of which the above are only examples - hence I find it problematic that the authors present the behaviour as evidence for one extremely specific model.

(3) Despite efforts to control confounds in the fMRI study, including two control experiments, I think some confounds remain.

For example, a network of regions is presented as correlating with the cumulative probability that there has been a regime shift in this block of 10 samples (Pt). However, regardless of the exact samples shown, doesn't Pt always increase with sample number (as by the time of later samples, there have been more opportunities for a regime shift)? Unless this is completely linear, the effect won't be controlled by including trial number as a co-regressor (which was done).

On the other hand, two additional fMRI experiments are done as control experiments and the effect of Pt in the main study is compared to Pt in these control experiments. Whilst I admire the effort in carrying out control studies, I can't understand how these particular experiment are useful controls. For example in experiment 3 participants simply type in numbers presented on the screen - how can we even have an estimate of Pt from this task?

(4) The Discussion is very long, and whilst a lot of related literature is cited, I found it hard to pin down within the discussion, what the key contributions of this study are. In my opinion it would be better to have a short but incisive discussion highlighting the advances in understanding that arise from the current study, rather than reviewing the field so broadly.

Editors' note: Reviewer #2 was unavailable to re-review the manuscript. Reviewer #3 was added for this round of review to ensure two reviewers and because of their expertise in the computational and modelling aspects of the work.

<https://doi.org/10.7554/eLife.104684.2.sa1>

Author response:

The following is the authors' response to the current reviews.

eLife Assessment

This study offers valuable insights into how humans detect and adapt to regime shifts, highlighting distinct contributions of the frontoparietal network and ventromedial prefrontal cortex to sensitivity to signal diagnosticity and transition probabilities. The combination of an innovative task design, behavioral modeling, and model-based fMRI analyses provides a solid foundation for the conclusions; however, the neuroimaging results have several limitations, particularly a potential confound between the posterior probability of a switch and the passage of time that may not be fully controlled by including trial number as a regressor. The control experiments intended to address this issue also appear conceptually inconsistent and, at the behavioral level, while informing participants of conditional probabilities rather than requiring learning is theoretically elegant, such information is difficult to apply accurately, as shown by well-documented challenges with conditional reasoning and base-rate neglect. Expressing these probabilities as natural frequencies rather than percentages may have improved comprehension. Overall, the study advances understanding of belief updating under uncertainty but would benefit from more intuitive probabilistic framing and stronger control of temporal confounds in future work.

We thank the editors for the assessment. The editor added several limitations based on the new reviewer 3 in this round, which we address below.

With regard to temporal confounds, we clarified in the main text and response to Reviewer 3 that we had already addressed the potential confound between posterior probability of a switch and passage of time in GLM-2 with the inclusion of intertemporal prior. After adding intertemporal prior in the GLM, we still observed the same fMRI results on probability estimates. In addition, we did two other robustness checks, which we mentioned in the manuscript.

With regard to response mode (probability estimation rather than choice or indicating natural frequencies), we wish to point out that the in previous research by Massey and Wu (2005), which the current study was based on, the concern of participants showing system-neglect tendencies due to the mode of information delivery, namely indicating beliefs through reporting probability estimates rather than through choice or other response mode was addressed. Massey and Wu (2005, Study 3) found the same biases when participants performed a choice task that did not require them to indicate probability estimates.

With regard to the control experiments, the control experiments in fact were not intended to address the confounds between posterior probability and passage of time. Rather, they aimed to address whether the neural findings were unique to change detection (Experiment 2) and to address visual and motor confounds (Experiment 3). These and the results of the control experiments were mentioned on page 18-19.

Finally, we wish to highlight that we had performed detailed model comparisons after reviewer 2's suggestions. Although reviewer 2 was unable to re-review the manuscript, we believe this provides insight into the literature on change detection. See "Incorporating signal dependency into system-neglect model led to better models for regime-shift detection" (p.27-30). The model comparison showed that system-neglect models that incorporate signal dependency are better models than the original system-neglect model in describing participants probability estimates. This suggests that people respond to change-consistent and change-inconsistent signals differently when judging whether the regime had changed. This was not reported in previous behavioral studies and was largely inspired by the neural

finding on signal dependency in the frontoparietal cortex. It indicates that neural findings can provide novel insights into computational modeling of behavior.

Public Reviews:

Reviewer #1 (Public review):

Summary:

The study examines human biases in a regime-change task, in which participants have to report the probability of a regime change in the face of noisy data. The behavioral results indicate that humans display systematic biases, in particular, overreaction in stable but noisy environments and underreaction in volatile settings with more certain signals. fMRI results suggest that a frontoparietal brain network is selectively involved in representing subjective sensitivity to noise, while the vmPFC selectively represents sensitivity to the rate of change.

Strengths:

- *The study relies on a task that measures regime-change detection primarily based on descriptive information about the noisiness and rate of change. This distinguishes the study from prior work using reversal-learning or change-point tasks in which participants are required to learn these parameters from experiences. The authors discuss these differences comprehensively.*
- *The study uses a simple Bayes-optimal model combined with model fitting, which seems to describe the data well. The model is comprehensively validated.*
- *The authors apply model-based fMRI analyses that provide a close link to behavioral results, offering an elegant way to examine individual biases.*

We thank the reviewer for the comments.

Weaknesses:

The authors have adequately addressed most of my prior concerns.

We thank the reviewer for recognizing our effort in addressing your concerns.

My only remaining comment concerns the z-test of the correlations. I agree with the non-parametric test based on bootstrapping at the subject level, providing evidence for significant differences in correlations within the left IFG and IPS.

However, the parametric test seems inadequate to me. The equation presented is described as the Fisher z-test, but the numerator uses the raw correlation coefficients (r) rather than the Fisher-transformed values (z). To my understanding, the subtraction should involve the Fisher z-scores, not the raw correlations.

More importantly, the Fisher z-test in its standard form assumes that the correlations come from independent samples, as reflected in the denominator (which uses the n of each independent sample). However, in my opinion, the two correlations are not independent but computed within-subject. In such cases, parametric tests should take into account the dependency. I believe one appropriate method for the current case (correlated correlation coefficients sharing a variable [behavioral slope]) is explained here:

Meng, X.-I., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. Psychological Bulletin, 111(1), 172-175. <https://doi.org/10.1037/0033-2909.111.1.172>

It should be implemented here:

Diedenhofen B, Musch J (2015) *cocor: A Comprehensive Solution for the Statistical Comparison of Correlations*. PLoS ONE 10(4): e0121945.
<https://doi.org/10.1371/journal.pone.0121945>

My recommendation is to verify whether my assumptions hold, and if so, perform a test that takes correlated correlations into account. Or, to focus exclusively on the non-parametric test.

In any case, I recommend a short discussion of these findings and how the authors interpret that some of the differences in correlations are not significant.

Thank you for the careful check. Yes. This was indeed a mistake from us. We also agree that the two correlations are not independent. Therefore, we modified the test that accounts for dependent correlations by following Meng et al. (1992) suggested by the reviewer.

We referred to the correlation between neural and behavioral sensitivity at change-consistent (blue) signals as r_{blue} , and that at change-inconsistent (red) signals as r_{red} . To statistically compare these two correlations, we adopted the approach of Meng et al. (1992), which specifically tests differences between dependent correlations according to the following equation

$$z = (z_{r1} - z_{r2}) \sqrt{\frac{N-3}{2(1-r_x^2)h}}$$

where N is the number of subjects, z_{r_i} is the Fisher z-transformed value of r_i , $r_1 = r_{blue}$ and $r_2 = r_{red}$. r_x is the correlation between the neural sensitivity at change-consistent signals and change-inconsistent signals.

$$h = \frac{1 - \bar{r}^2}{1 - r_x^2} = 1 + \frac{\bar{r}^2}{1 - r_x^2} (1 - f)$$

$$f = \frac{1 - r_x}{2(1 - \bar{r}^2)}, \text{ which must be } \leq 1$$

Where $\bar{r}^2$ is the mean of the r_i^2 , and f should be set to 1 if $\bar{r}^2 > 1$.

We found that among the five ROIs in the frontoparietal network, two of them, namely the left IFG and left IPS, the difference in correlation was significant (one-tailed z test; left IFG: $z = 1.8908$, $p = 0.0293$; left IPS: $z = 2.2584$, $p = 0.0049$). For the remaining three ROIs, the difference in correlation was not significant (dmPFC: $z = 0.9522$, $p = 0.1705$; right IFG: $z = 0.9860$, $p = 0.1621$; right IPS: $z = 1.4833$, $p = 0.0690$). We chose one-tailed test because we already know the correlation under the blue signals was significantly greater than 0. These updated results are consistent with the nonparametric tests we had already performed and we will update them in the revised manuscript.

Reviewer #3 (Public review):

This study concerns how observers (human participants) detect changes in the statistics of their environment, termed regime shifts. To make this concrete, a series of 10 balls are drawn from an urn that contains mainly red or mainly blue balls. If there is a regime shift, the urn is changed over (from mainly red to mainly blue) at some point in the 10 trials. Participants report their belief that there has been a regime shift as a % probability. Their judgement should (mathematically) depend on the prior probability of a regime shift (which is set at one of three levels) and the strength of evidence (also one of three levels, operationalized as the proportion of red balls in the mostly-blue urn and vice versa). Participants are directly instructed of the prior probability of regime shift and

proportion of red balls, which are presented on-screen as numerical probabilities. The task therefore differs from most previous work on this question in that probabilities are instructed rather than learned by observation, and beliefs are reported as numerical probabilities rather than being inferred from participants' choice behaviour (as in many bandit tasks, such as Behrens 2007 Nature Neurosci).

The key behavioural finding is that participants over-estimate the prior probability of regime change when it is low, and under estimate it when it is high; and participants over-estimate the strength of evidence when it is low and under-estimate it when it is high. In other words participants make much less distinction between the different generative environments than an optimal observer would. This is termed 'system neglect'. A neuroeconomic-style mathematical model is presented and fit to data.

Functional MRI results show that strength of evidence for a regime shift (roughly, the surprise associated with a blue ball from an apparently red urn) is associated with activity in the frontal-parietal orienting network. Meanwhile, at time-points where the probability of a regime shift is high, there is activity in another network including vmPFC. Both networks show individual differences effects, such that people who were more sensitive to strength of evidence and prior probability show more activity in the frontal-parietal and vmPFC-linked networks respectively.

We thank the reviewer for the overall descriptions of the manuscript.

Strengths:

(1) The study provides a different task for looking at change-detection and how this depends on estimates of environmental volatility and sensory evidence strength, in which participants are directly and precisely informed of the environmental volatility and sensory evidence strength rather than inferring them through observation as in most previous studies

(2) Participants directly provide belief estimates as probabilities rather than experimenters inferring them from choice behaviour as in most previous studies

(3) The results are consistent with well-established findings that surprising sensory events activate the frontal-parietal orienting network whilst updating of beliefs about the word ('regime shift') activates vmPFC.

Thank you for these assessments.

Weaknesses:

(1) The use of numerical probabilities (both to describe the environments to participants, and for participants to report their beliefs) may be problematic because people are notoriously bad at interpreting probabilities presented in this way, and show poor ability to reason with this information (see Kahneman's classic work on probabilistic reasoning, and how it can be improved by using natural frequencies). Therefore the fact that, in the present study, people do not fully use this information, or use it inaccurately, may reflect the mode of information delivery.

We appreciate the reviewer's concern on this issue. The concern was addressed in Massey and Wu (2005) as participants performed a choice task in which they were not asked to provide probability estimates (Study 3 in Massey and Wu, 2005). Instead, participants in Study 3 were asked to predict the color of the ball before seeing a signal. This was a more intuitive way of indicating his or her belief about regime shift. The results from the choice task were identical to those found in the probability estimation task (Study 1 in Massey and Wu). We

take this as evidence that the system-neglect behavior the participants showed was less likely to be due to the mode of information delivery.

(2) Although a very precise model of 'system neglect' is presented, many other models could fit the data.

For example, you would get similar effects due to attraction of parameter estimates towards a global mean - essentially application of a hyper-prior in which the parameters applied by each participant in each block are attracted towards the experiment-wise mean values of these parameters. For example, the prior probability of regime shift ground-truth values [0.01, 0.05, 0.10] are mapped to subjective values of [0.037, 0.052, 0.069]; this would occur if observers apply a hyper-prior that the probability of regime shift is about 0.05 (the average value over all blocks). This 'attraction to the mean' is a well-established phenomenon and cannot be ruled out with the current data (I suppose you could rule it out by comparing to another dataset in which the mean ground-truth value was different).

We thank the reviewer for this comment. It is true that the system-neglect model is not entirely inconsistent with regression to the mean, regardless of whether the implementation has a hyper prior or not. In fact, our behavioral measure of sensitivity to transition probability and signal diagnosticity, which we termed the behavioral slope, is based on linear regression analysis. In general, the modeling approach in this paper is to start from a generative model that defines ideal performance and consider modifying the generative model when systematic deviations in actual performance from the ideal is observed. In this approach, a generative model with hyper-prior would be more complex to begin with, and a regression to the mean idea by itself does not generate a priori predictions.

More generally, any model in which participants don't fully use the numerical information they were given would produce apparent 'system neglect'. Four qualitatively different example reasons are: 1. Some individual participants completely ignored the probability values given. 2. Participants did not ignore the probability values given, but combined them with a hyperprior as above. 3. Participants had a reporting bias where their reported beliefs that a regime-change had occurred tend to be shifted towards 50% (rather than reporting 'confident' values such 5% or 95%). 4. Participants underweighted probability outliers resulting in underweighting of evidence in the 'high signal diagnosticity' environment (10.1016/j.neuron.2014.01.020)

In summary I agree that any model that fits the data would have to capture the idea that participants don't differentiate between the different environments as much as they should, but I think there are a number of qualitatively different reasons why they might do this - of which the above are only examples - hence I find it problematic that the authors present the behaviour as evidence for one extremely specific model.

Thank you for raising this point. The modeling principle we adopt is the following. We start from the normative model—the Bayesian model—that defined what normative behavior should look like. We compared participants' behavior with the Bayesian model and found systematic deviations from it. To explain those systematic deviations, we considered modeling options within the confines of the same modeling framework. In other words, we considered a parameterized version of the Bayesian model, which is the system-neglect model and examined through model comparison the best modeling choice. This modeling approach is not uncommon, and many would agree this is the standard approach in economics and psychology. For example, Kahneman and Tversky adopted this approach when proposing prospect theory, a modification of expected utility theory where expected utility theory can be seen as one specific model for how utility of an option should be computed.

(3) Despite efforts to control confounds in the fMRI study, including two control experiments, I think some confounds remain.

For example, a network of regions is presented as correlating with the cumulative probability that there has been a regime shift in this block of 10 samples (Pt). However, regardless of the exact samples shown, doesn't Pt always increase with sample number (as by the time of later samples, there have been more opportunities for a regime shift)? Unless this is completely linear, the effect won't be controlled by including trial number as a co-regressor (which was done).

Thank you for raising this concern. Yes, Pt always increases with sample number regardless of evidence (seeing change-consistent or change-inconsistent signals). This is captured by the 'intertemporal prior' in the Bayesian model, which we included as a regressor in our GLM analysis (GLM-2), in addition to Pt. In short, GLM-1 had Pt and sample number. GLM-2 had Pt, intertemporal prior, and sample number, among other regressors. And we found that, in both GLM-1 and GLM-2, both vmPFC and ventral striatum correlated with Pt.

To make this clearer, we updated the main text to further clarify this on p.18:

On the other hand, two additional fMRI experiments are done as control experiments and the effect of Pt in the main study is compared to Pt in these control experiments. Whilst I admire the effort in carrying out control studies, I can't understand how these particular experiment are useful controls. For example in experiment 3 participants simply type in numbers presented on the screen - how can we even have an estimate of Pt from this task?

We thank the reviewer for this comment. The purpose of Experiment 3 was to control for visual and motor confounds. In other words, if subjects saw the similar visual layout and were just instructed to press numbers, would we observe the vmPFC, ventral striatum, and the frontoparietal network like what we did in the main experiment (Experiment 1)?

The purpose of Experiment 2 was to establish whether what we found about Pt was unique to change detection. In Experiment 2, subjects estimated the probability that the current regime is the blue regime (just as they did in Experiment 1) except that there were no regime shifts involved. In other words, it is possible that the regions we identified were generally associated with probability estimation and not particularly about change detection. And we used Experiment 2 to examine whether this were true.

(4) The Discussion is very long, and whilst a lot of related literature is cited, I found it hard to pin down within the discussion, what the key contributions of this study are. In my opinion it would be better to have a short but incisive discussion highlighting the advances in understanding that arise from the current study, rather than reviewing the field so broadly.

Thank you. We received different feedbacks from previous reviews on what to include in Discussion. To address the reviewer's concern, we will revise the Discussion to better highlight the key contributions of the current study at the beginning of Discussion.

Recommendations for the authors:

Reviewer #3 (Recommendations for the authors):

Many of the figures are too tiny - the writing is very small, as are the pictures of brains. I'd suggest adjusting these so they will be readable without enlarging.

Thank you. We will enlarge the figures to make them more readable.

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public review):

Summary:

The study examines human biases in a regime-change task, in which participants have to report the probability of a regime change in the face of noisy data. The behavioral results indicate that humans display systematic biases, in particular, overreaction in stable but noisy environments and underreaction in volatile settings with more certain signals. fMRI results suggest that a frontoparietal brain network is selectively involved in representing subjective sensitivity to noise, while the vmPFC selectively represents sensitivity to the rate of change.

Strengths:

(1) The study relies on a task that measures regime-change detection primarily based on descriptive information about the noisiness and rate of change. This distinguishes the study from prior work using reversal-learning or change-point tasks in which participants are required to learn these parameters from experiences. The authors discuss these differences comprehensively.

Thank you for recognizing our contribution to the regime-change detection literature and our effort in discussing our findings in relation to the experience-based paradigms.

(2) The study uses a simple Bayes-optimal model combined with model fitting, which seems to describe the data well.

Thank you for recognizing the contribution of our Bayesian framework and system neglect model.

(3) The authors apply model-based fMRI analyses that provide a close link to behavioral results, offering an elegant way to examine individual biases.

Thank you for recognizing our execution of model-based fMRI analyses and effort in using those analyses to link with behavioral biases.

Weaknesses:

My major concern is about the correlational analysis in the section "Under- and overreactions are associated with selectivity and sensitivity of neural responses to system parameters", shown in Figures 5c and d (and similarly in Figure 6). The authors argue that a frontoparietal network selectively represents sensitivity to signal diagnosticity, while the vmPFC selectively represents transition probabilities. This claim is based on separate correlational analyses for red and blue across different brain areas. The authors interpret the finding of a significant correlation in one case (blue) and an insignificant correlation (red) as evidence of a difference in correlations (between blue and red) but don't test this directly. This has been referred to as the "interaction fallacy" (Nieuwenhuis et al., 2011; Makin & Orban de Xivry 2019). Not directly testing the difference in correlations (but only the differences to zero for each case) can lead to wrong conclusions. For example, in Figure 5c, the correlation for red is $r = 0.32$ (not significantly different from zero) and $r = 0.48$ (different from zero). However, the difference between the two is 0.1, and it is likely that this difference itself is not significant. From a statistical perspective, this corresponds to an interaction effect that has to be tested directly. It is my understanding that analyses in Figure 6 follow the same approach.

Relevant literature on this point is:

Nieuwenhuis, S, Forstmann, B & Wagenmakers, EJ (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat Neurosci* 14, 11051107. <https://doi.org/10.1038/nn.2886>

Makin TR, Orban de Xivry, JJ (2019). *Science Forum: Ten common statistical mistakes to watch out for when writing or reviewing a manuscript.* *eLife* 8:e48175. <https://doi.org/10.7554/eLife.48175>

There is also a blog post on simulation-based comparisons, which the authors could check out: <https://garstats.wordpress.com/2017/03/01/comp2dcorr/>

I recommend that the authors carefully consider what approach works best for their purposes. It is sometimes recommended to directly compare correlations based on Monte-Carlo simulations (cf Makin & Orban). It might also be appropriate to run a regression with the dependent variable brain activity (Y) and predictors brain area (X) and the model-based term of interest (Z). In this case, they could include an interaction term in the model:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X \cdot Z$$

The interaction term reflects if the relationship between the model term Z and brain activity Y is conditional on the brain area of interest X.

Thank you for the suggestion. In response, we tested for the difference in correlation both parametrically and nonparametrically. The results were identical. In the parametric test, we used the Fisher z transformation to transform the difference in correlation coefficients to the z statistic. That is, for two correlation coefficients, r_1 (with sample size n_1) and r_2 (with sample size n_2), the z statistic of the difference in correlation is given by

$$z = \frac{r_1 - r_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

We referred to the correlation between neural and behavioral sensitivity at change-consistent (blue) signals as r_{blue} , and that at change-inconsistent (red) signals as r_{red} . For the Fisher z transformation $r_1 = r_{blue}$ and $r_2 = r_{red}$. We found that among the five ROIs in the frontoparietal network, two of them, namely the left IFG and left IPS, the difference in correlation was significant (one-tailed z test; left IFG: $z = 1.8355$, $p = 0.0332$; left IPS: $z = 2.3782$, $p = 0.0087$). For the remaining three ROIs, the difference in correlation was not significant (dmPFC: $z = 0.7594$, $p = 0.2238$; right IFG: $z = 0.9068$, $p = 0.1822$; right IPS: $z = 1.3764$, $p = 0.0843$). We chose one-tailed test because we already know the correlation under the blue signals was significantly greater than 0.

In the nonparametric test, we performed nonparametric bootstrapping to test for the difference in correlation (Efron & Tibshirani, 1994). We resampled with replacement the dataset (subject-wise) and used the resampled dataset to compute the difference in correlation. We then repeated the above for 100,000 times so as to estimate the distribution of the difference in correlation coefficients, tested for significance and estimated p-value based on this distribution. Consistent with our parametric tests, here we also found that the difference in correlation was significant in left IFG and left IPS (left IFG: $r_{blue} - r_{red} = 0.46$, $p = 0.0496$; left IPS: $r_{blue} - r_{red} = 0.5306$, $p = 0.0041$), but was not significant in dmPFC, right IFG, and right IPS (dmPFC: $r_{blue} - r_{red} = 0.1634$, $p = 0.1919$; right IFG: $r_{blue} - r_{red} = 0.2123$, $p = 0.1681$; right IPS: $r_{blue} - r_{red} = 0.3434$, $p = 0.0631$).

In summary, we found that neural sensitivity to signal diagnosticity in the frontoparietal network measured at change-consistent signals significantly correlated with individual

subjects' behavioral sensitivity to signal diagnosticity (r_{blue}). By contrast, neural sensitivity to signal diagnosticity measured at change-inconsistent did not significantly correlate with behavioral sensitivity (r_{red}). The difference in correlation, $r_{blue} - r_{red}$, however, was statistically significant in some (left IPS and left IFG) but not all brain regions within the frontoparietal network.

To incorporate these updates, we added descriptions of the methods and results in the revised manuscript. In the Results section (p.26-27):

"We further tested, for each brain region, whether the difference in correlation was significant using both parametric and nonparametric tests (see Parametric and nonparametric tests for difference in correlation coefficients in Methods). The results were identical. In the parametric test, we used the Fisher z transformation to transform the difference in correlation coefficients to the z statistic. We found that among the five ROIs in the frontoparietal network, two of them, namely the left IFG and left IPS, the difference in correlation was significant (one-tailed z test; left IFG: $z = 1.8355$, $p = 0.0332$; left IPS: $z = 2.3782$, $p = 0.0087$). For the remaining three ROIs, the difference in correlation was not significant (dmPFC: $z = 0.7594$, $p = 0.2238$; right IFG: $z = 0.9068$, $p = 0.1822$; right IPS: $z = 1.3764$, $p = 0.0843$). We chose one-tailed test because we already know the correlation under change-consistent signals was significantly greater than 0. In the nonparametric test, we performed nonparametric bootstrapping to test for the difference in correlation. We referred to the correlation between neural and behavioral sensitivity at change-consistent (blue) signals as r_{blue} , and that at change-inconsistent (red) signals as r_{red} . Consistent with the parametric tests, we also found that the difference in correlation was significant in left IFG and left IPS (left IFG: $r_{blue} - r_{red} = 0.46$, $p = 0.0496$; left IPS: $r_{blue} - r_{red} = 0.5306$, $p = 0.0041$), but was not significant in dmPFC, right IFG, and right IPS (dmPFC: $r_{blue} - r_{red} = 0.1634$, $p = 0.1919$; right IFG: $r_{blue} - r_{red} = 0.2123$, $p = 0.1681$; right IPS: $r_{blue} - r_{red} = 0.3434$, $p = 0.0631$). In summary, we found that neural sensitivity to signal diagnosticity measured at change-consistent signals significantly correlated with individual subjects' behavioral sensitivity to signal diagnosticity. By contrast, neural sensitivity to signal diagnosticity measured at change-inconsistent signals did not significantly correlate with behavioral sensitivity. The difference in correlation, however, was statistically significant in some (left IPS and left IFG) but not all brain regions within the frontoparietal network."

In the Methods section, we added on p.53:

"Parametric and nonparametric tests for difference in correlation coefficients. We implemented both parametric and nonparametric tests to examine whether the difference in Pearson correlation coefficients was significant. In the parametric test, we used the Fisher z transformation to transform the difference in correlation coefficients to the z statistic. That is, for two correlation coefficients, r_1 (with sample size n_2) and r_2 , (with sample size n_1), the z statistic of the difference in correlation is given by

$$z = \frac{r_1 - r_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

We referred to the correlation between neural and behavioral sensitivity at changeconsistent (blue balls) signals as r_{blue} , and that at change-inconsistent (red balls) signals as r_{red} . For the Fisher z transformation, $r_1 = r_{blue}$ and $r_2 = r_{red}$. In the nonparametric test, we performed nonparametric bootstrapping to test for the difference in correlation (Efron & Tibshirani, 1994). That is, we resampled with replacement the dataset (subject-wise) and used the resampled dataset to compute the difference in correlation. We then repeated the above for 100,000 times so as to estimate the distribution of the difference in correlation coefficients, tested for significance and estimated p-value based on this distribution."

Another potential concern is that some important details about the parameter estimation for the system-neglect model are missing. In the respective section in the methods, the

authors mention a nonlinear regression using Matlab's "fitnlm" function, but it remains unclear how the model was parameterized exactly. In particular, what are the properties of this nonlinear function, and what are the assumptions about the subject's motor noise? I could imagine that by using the inbuild function, the assumption was that residuals are Gaussian and homoscedastic, but it is possible that the assumption of homoscedasticity is violated, and residuals are systematically larger around $p=0.5$ compared to $p=0$ and $p=1$. Relatedly, in the parameter recovery analyses, the authors assume different levels of motor noise. Are these values representative of empirical values?

We thank the reviewer for this excellent point. The reviewer touched on model parameterization, assumption of noise, and parameter recovery analysis. We answered these questions point-by-point below.

On how our model was parameterized

We parameterized the model according to the system-neglect model in Eq. (2) and estimated the alpha parameter separately for each level of transition probability and the beta parameter separately for each level of signal diagnosticity. As a result, we had a total of 6 parameters (3 alpha and 3 beta parameters) in the model. The system-neglect model is then called by fitnlm so that these parameters can be estimated. The term 'nonlinear' regression in fitnlm refers to the fact that you can specify any model (in our case the system-neglect model) and estimate its parameters when calling this function. In our use of fitnlm, we assume that the noise is Gaussian and homoscedastic (the default option).

On the assumptions about subject's motor noise

We actually never called the noise 'motor' because it can be estimation noise as well. In the context of fitnlm, we assume that the noise is Gaussian and homoscedastic.

On the possibility that homoscedasticity is violated

We take the reviewer's point. In response, we separately estimated the residual standard deviation at different probability intervals ([0.0–0.2), [0.2–0.4), [0.4–0.6), [0.6–0.8), and [0.8–1.0]). The result is shown in the figure below. The black data points are the average residual standard deviation (across subjects) and the error bars are the standard error of the mean. The residual standard deviation is indeed heteroscedastic—smallest at 0.1 probability and increasing as probability increases and asymptote at 0.5 (Fig. S4).

To examine how this would affect model fitting (parameter estimation), we performed parameter recovery analysis based on these empirically estimated, probability-dependent residual standard deviation. That is, we simulated subjects' probability estimates using the system-neglect model and added the heteroscedastic noise according to the empirical values and then estimated the parameter estimates of the system-neglect model. The recovered parameter estimates did not seem to be affected by the heteroscedasticity of the variance. The parameter recovery results were identical to the parameter recovery results when homoscedasticity was assumed. This suggested that although homoscedasticity was violated, it did not affect the accuracy of the parameter estimates (Fig.S4).

We added a section 'Impact of noise homoscedasticity on parameter estimation' in Methods section (p.47-48) and a figure in the supplement (Fig. S4) to describe this:

On whether the noise levels in parameter recovery analysis are representative of empirical values

To address the reviewer's question, we conducted a new analysis using maximum likelihood estimation to simultaneously estimate the system-neglect model and the noise level of each

individual subject. To estimate each subject's noise level, we incorporated a noise parameter into the system-neglect model. We assumed that probability estimates are noisy and modeled them with a Gaussian distribution where the noise parameter (σ , $-$, $\&$) is the standard deviation. At each period, a probability estimate of regime shift was computed according to the system-neglect model where Θ is the set of parameters including parameters in the system-neglect model and the noise parameter. The likelihood function, $L(\Theta)$, is the probability of observing the subject's actual probability estimate at period t , p , given Θ , $L(\Theta) = P(p | \Theta)$. Since we modeled the noisy probability estimates with a Gaussian distribution, we can therefore express $L(\Theta)$ as $L(\Theta) \sim N(p; p^{*+}, \sigma, -, \&)$ where p^{*+} is the probability estimate predicted by the system-neglect (SN) model at period t . As a reminder, we referred to a 'period' as the time when a new signal appeared during a trial (for a given transition probability and signal diagnosticity). To find that maximum likelihood estimates of Θ MLE, we summed over all periods the negative natural logarithm of likelihood and used MATLAB's `fmincon` function to find Θ MLE. Across subjects, we found that the mean noise estimate was 0.1735 and ranged from 0.1118 to 0.2704 (Supplementary Figure S3)."

Compared with our original parameter recovery analysis where the maximum noise level was set at 0.1, our data indicated that some subjects' noise was larger than this value. Therefore, we expanded our parameter recovery analysis to include noise levels beyond 0.1 to up to 0.3. The results are now updated in Supplementary Fig. S3.

We updated the parameter recovery section (p. 47) in Methods:

The main study is based on N=30 subjects, as are the two control studies. Since this work is about individual differences (in particular w.r.t. to neural representations of noise and transition probabilities in the frontoparietal network and the vmPFC), I'm wondering how robust the results are. Is it likely that the results would replicate with a larger number of subjects? Can the two control studies be leveraged to address this concern to some extent?

We can address the issue of robustness through looking at the effect size. In particular, with respect to individual differences in neural sensitivity of transition probability and signal diagnosticity, since the significant correlation coefficients between neural and behavioral sensitivity were between 0.4 and 0.58 for signal diagnosticity in frontoparietal network (Fig. 5C), and -0.38 and -0.37 for transition probability in vmPFC (Fig. 5D), the effect size of these correlation coefficients was considered medium to large (Cohen, 1992).

It would be challenging to use the control studies to address the robustness concern. The two control studies did not allow us to examine individual differences – in particular with respect to neural selectivity of noise and transition probability – and therefore we think it is less likely to leverage the control studies. Having said that, it is possible to look at neural selectivity of noise (signal diagnosticity) in the first control experiment where subjects estimated the probability of blue regime in a task where there was no regime change (transition probability was 0). However, the fact that there were no regime shifts changed the nature of the task. Instead of always starting at the Red regime in the main experiment, in the first control experiment we randomly picked the regime to draw the signals from. It also changed the meaning and the dynamics of the signals (red and blue) that would appear. In the main experiment the blue signal is a signal consistent with change, but in the control experiment this is no longer the case. In the main experiment, the frequency of blue signals is contingent upon both noise and transition probability. In general, blue signals are less frequent than red signals because of small transition probabilities. But in the first control experiment, the frequency of blue signals may not be less frequent because the regime was blue in half of the trials. Due to these differences, we do not see how analyzing the control experiments could help in establishing robustness because we do not have a good prediction as to whether and how the neural selectivity would be impacted by these differences.

It seems that the authors have not counterbalanced the colors and that subjects always reported the probability of the blue regime. If so, I'm wondering why this was not counterbalanced.

We are aware of the reviewer's concern. The first reason we did not do these (color counterbalancing and report blue/red regime balancing) was to not confuse the subjects in an already complicated task. Balancing these two variables also comes at the cost of sample size, which was the second reason we did not do it. Although we can elect to do these balancing at the between-subject level to not impact the task complexity, we could have introduced another confound that is the individual differences in how people respond to these variables. This is the third reason we were hesitant to do these counterbalancing.

Reviewer #2 (Public review):

Summary:

This paper focuses on understanding the behavioral and neural basis of regime shift detection, a common yet hard problem that people encounter in an uncertain world.

Using a regime-shift task, the authors examined cognitive factors influencing belief updates by manipulating signal diagnosticity and environmental volatility. Behaviorally, they have found that people demonstrate both over and under-reaction to changes given different combinations of task parameters, which can be explained by a unified system-neglect account. Neurally, the authors have found that the vmPFC-striatum network represents current belief as well as belief revision unique to the regime detection task. Meanwhile, the frontoparietal network represents cognitive factors influencing regime detection i.e., the strength of the evidence in support of the regime shift and the intertemporal belief probability. The authors further link behavioral signatures of system neglect with neural signals and have found dissociable patterns, with the frontoparietal network representing sensitivity to signal diagnosticity when the observation is consistent with regime shift and vmPFC representing environmental volatility, respectively. Together, these results shed light on the neural basis of regime shift detection especially the neural correlates of bias in belief update that can be observed behaviorally.

Strengths:

(1) The regime-shift detection task offers a solid ground to examine regime-shift detection without the potential confounding impact of learning and reward. Relatedly, the system-neglect modeling framework provides a unified account for both over or under-reacting to environmental changes, allowing researchers to extract a single parameter reflecting people's sensitivity to changes in decision variables and making it desirable for neuroimaging analysis to locate corresponding neural signals.

Thank you for recognizing our task design and our system-neglect computational framework in understanding change detection.

(2) The analysis for locating brain regions related to belief revision is solid. Within the current task, the authors look for brain regions whose activation covary with both current belief and belief change. Furthermore, the authors have ruled out the possibility of representing mere current belief or motor signal by comparing the current study results with two other studies. This set of analyses is very convincing.

Thank you for recognizing our control studies in ruling out potential motor confounds in our neural findings on belief revision.

(3) The section on using neuroimaging findings (i.e., the frontoparietal network is sensitive to evidence that signals regime shift) to reveal nuances in behavioral data (i.e.,

belief revision is more sensitive to evidence consistent with change) is very intriguing. I like how the authors structure the flow of the results, offering this as an extra piece of behavioral findings instead of ad-hoc implanting that into the computational modeling.

Thank you for appreciating how we showed that neural insights can lead to new behavioral findings.

Weaknesses:

(1) The authors have presented two sets of neuroimaging results, and it is unclear to me how to reason between these two sets of results, especially for the frontoparietal network. On one hand, the frontoparietal network represents belief revision but not variables influencing belief revision (i.e., signal diagnosticity and environmental volatility). On the other hand, when it comes to understanding individual differences in regime detection, the frontoparietal network is associated with sensitivity to change and consistent evidence strength. I understand that belief revision correlates with sensitivity to signals, but it can probably benefit from formally discussing and connecting these two sets of results in discussion. Relatedly, the whole section on behavioral vs. neural slope results was not sufficiently discussed and connected to the existing literature in the discussion section. For example, the authors could provide more context to reason through the finding that striatum (but not vmPFC) is not sensitive to volatility.

We thank the reviewer for the valuable suggestions.

With regard to the first comment, we wish to clarify that we did not find frontoparietal network to represent belief revision. It was the vmPFC and ventral striatum that we found to represent belief revision (delta Pt in Fig. 3). For the frontoparietal network, we identified its involvement in our task through finding that its activity correlated with strength of change evidence (Fig. 4) and individual subjects' sensitivity to signal diagnosticity (Fig. 5). Conceptually, these two findings reflect how individuals interpret the signals (signals consistent or inconsistent with change) in light of signal diagnosticity. This is because (1) strength of change evidence is defined as signals (+1 for signal consistent with change, and -1 for signal inconsistent with change) multiplied by signal diagnosticity and (2) sensitivity to signal diagnosticity reflects how individuals subjectively evaluate signal diagnosticity. At the theoretical level, these two findings can be interpreted through our computational framework in that both the strength of change evidence and sensitivity to signal diagnosticity contribute to estimating the likelihood of change (Eqs. 1 and 2). We added a paragraph in Discussion to talk about this.

We added on p. 36:

“For the frontoparietal network, we identified its involvement in our task through finding that its activity correlated with strength of change evidence (Fig. 4) and individual subjects' sensitivity to signal diagnosticity (Fig. 5). Conceptually, these two findings reflect how individuals interpret the signals (signals consistent or inconsistent with change) in light of signal diagnosticity. This is because (1) strength of change evidence is defined as signals (+1 for signal consistent with change, and -1 for signal inconsistent with change) multiplied by signal diagnosticity and (2) sensitivity to signal diagnosticity reflects how individuals subjectively evaluate signal diagnosticity. At the theoretical level, these two findings can be interpreted through our computational framework in that both the strength of change evidence and sensitivity to signal diagnosticity contribute to estimating the likelihood of change (Equations 1 and 2 in Methods).”

With regard to the second comment, we added a discussion on the behavioral and neural slope comparison. We pointed out previous papers conducting similar analysis (Vilares et al., 2011; Ting et al., 2015; Yang & Wu, 2020), their findings and how they relate to our results. Vilares et al. found that sensitivity to prior information (uncertainty in prior distribution) in

the orbitofrontal cortex (OFC) and putamen correlated with behavioral measure of sensitivity to prior. In the current study, transition probability acts as prior in the system-neglect framework (Eq. 1) and we found that ventromedial prefrontal cortex represents subjects' sensitivity to transition probability. Together, these results suggest that OFC (with vmPFC being part of OFC, see Wallis, 2011) is involved in the subjective evaluation of prior information in both static (Vilares et al., 2011) and dynamic environments (current study).

We added on p. 37-38:

“In the current study, our psychometric-neurometric analysis focused on comparing behavioral sensitivity with neural sensitivity to the system parameters (transition probability and signal diagnosticity). We measured sensitivity by estimating the slope of behavioral data (behavioral slope) and neural data (neural slope) in response to the system parameters. Previous studies had adopted a similar approach (Ting et al., 2015a; Vilares et al., 2012; Yang & Wu, 2020). For example, Vilares et al. (2012) found that sensitivity to prior information (uncertainty in prior distribution) in the orbitofrontal cortex (OFC) and putamen correlated with behavioral measure of sensitivity to the prior.

In the current study, transition probability acts as prior in the system-neglect framework (Eq. 2 in Methods) and we found that ventromedial prefrontal cortex represents subjects' sensitivity to transition probability. Together, these results suggest that OFC (with vmPFC being part of OFC, see Wallis, 2011) is involved in the subjective evaluation of prior information in both static (Vilares et al., 2012) and dynamic environments (current study). In addition, distinct from vmPFC in representing sensitivity to transition probability or prior, we found through the behavioral-neural slope comparison that the frontoparietal network represents how sensitive individual decision makers are to the diagnosticity of signals in revealing the true state (regime) of the environment.”

(2) More details are needed for behavioral modeling under the system-neglect framework, particularly results on model comparison. I understand that this model has been validated in previous publications, but it is unclear to me whether it provides a superior model fit in the current dataset compared to other models (e.g., a model without α or β). Relatedly, I wonder whether the final result section can be incorporated into modeling as well - i.e., the authors could test a variant of the model with two β s depending on whether the observation is consistent with a regime shift and conduct model comparison.

Thank you for the great suggestion. We rewrote the final Results section to specifically focus on model comparison. To address the reviewer's suggestion (separately estimate beta parameters for change-consistent and change-inconsistent signals), we indeed found that these models were better than the original system-neglect model.

To incorporate these new findings, we rewrote the entire final result section “Incorporating signal dependency into system-neglect model led to better models for regime-shift detection” (p.28-30).

Recommendations for the authors:

Reviewer #1 (Recommendations for the authors):

(1) Use line numbers for the next round of reviews.

We added line numbers in the revised manuscript.

(2) Figure 2b: Can the empirical results be reproduced by the system-neglect model? This would complement the analyses presented in Figure S4.

Yes. We now add Figure S6 based on system-neglect model fits. For each subject, we first computed period-by-period probability estimates based on the parameter estimates of the system-neglect model. Second, we computed index of overreaction (IO) for each combination of transition probability and signal diagnosticity. Third, we plot the IO like we did using empirical results in Fig. 2b. We found that the empirical results in Fig. 2b are similar to the system-neglect model shown in Figure S6, indicating that the empirical results can be reproduced by the model.

(3) Page 14: Instead of referring to the "Methods" in general, you could be more specific about where the relevant information can be found.

Fixed. We changed "See Methods" to "See System-neglect model in Methods".

(4) Page 18: Consider avoiding the term "more significantly". Consider effect sizes if interested in comparing effects to each other.

Fixed. On page 19, we changed that to

"In the second analysis, we found that for both vmPFC and ventral striatum, the regression coefficient of P was significantly different between Experiment 1 and Experiment 2 (Fig. 3C) and between Experiment 1 and Experiment 3 (Fig. 3D; also see Tables S5 and S6 in SI)."

(5) Page 30: Cite key studies using reversal-learning paradigms. Currently, readers less familiar with the literature might have difficulties with this.

We now cite key studies using reversal-learning paradigms on p.32:

"Our work is closely related to the reversal-learning paradigm—the standard paradigm in neuroscience and psychology to study change detection (Fellows & Farah, 2003; Izquierdo et al., 2017; O'Doherty et al., 2001; Schoenbaum et al., 2000; Walton et al., 2010). In a typical reversal-learning task, human or animal subjects choose between two options that differ in the reward magnitude or probability of receiving a reward. Through reward feedback the participants gradually learn the reward contingencies associated with the options and have to update knowledge about reward contingencies when contingencies are switched in order to maximize rewards."

Reviewer #2 (Recommendations for the authors):

*(1) Some literature on change detection seems missing. For example, the author should also cite Muller, T. H., Mars, R. B., Behrens, T. E., & O'Reilly, J. X. (2019). Control of entropy in neural models of environmental state. *elife*, 8, e39404. This paper suggests that medial PFC is correlated with the entropy of the current state, which is closely related to regime change and environmental volatility.*

Thank you for pointing to this paper. We have now added it and other related papers in the Introduction and Discussion.

In Introduction, we added on p.5-6:

"Different behavioral paradigms, most notably reversal learning, and computational models were developed to investigate its neurocomputational substrates (Behrens et al., 2007; Izquierdo et al., 2017; Payzan-LeNestour et al., 2011, 2013; Nasser et al., 2010; McGuire et al., 2014; Muller et al., 2019). Key findings on the neural implementations for such learning include identifying brain areas and networks that track volatility in the environment (rate of change) (Behrens et al., 2007), the uncertainty or entropy of the current state of the environment (Muller et al., 2019), participants' beliefs about change (Payzan-LeNestour et al.,

2011; McGuire et al., 2014; Kao et al., 2020), and their uncertainty about whether a change had occurred (McGuire et al., 2014; Kao et al., 2020).”

In Discussion (p.35), we added a new paragraph:

“Related to OFC function in decision making and reinforcement learning, Wilson et al. (2014) proposed that OFC is involved in inferring the current state of the environment. For example, medial OFC had been shown to represent probability distribution on possible states of the environment (Chan et al., 2016), the current task state (Schuck et al., 2016) and uncertainty or entropy associated with the state of the environment (Muller et al., 2019). In the context of regime-shift detection, regimes can be regarded as states of the environment and therefore a change in regime indicates a change in the state of the environment. Muller et al. (2019) found that in dynamic environments where changes in the state of the environment happen regularly, medial OFC represented the level of uncertainty in the current state of the environment. Our finding that vmPFC represented individual participants’ probability estimates of regime shifts suggest that vmPFC and/or OFC are involved in inferring the current state of the environment through estimating whether the state has changed. Our finding that vmPFC represented individual participants’ sensitivity to transition probability further suggest that vmPFC and/or OFC contribute to individual participants’ biases in state inference (over- and underreactions to change) in how these brain areas respond to the volatility of the environment.”

(2) The language used when describing the selective relationship between frontoparietal network activation and change-consistent signal can be clearer. When describing separating those two signals, the authors refer to them as when the 'blue' signal shows up and when the 'red' signal shows up, assuming that the current belief state is blue. This is a little confusing cuz it is hard to keep in mind what is the default color in this example. It would be more intuitive if the author used language such as the 'change consistent' signal.

Thank you for the suggestion. We have changed the wording according to your suggestion. That is, we say ‘change-consistent (blue) signals’ and ‘change-inconsistent (red) signals’ throughout pages 22-28.

(3) Figure 4B highlights dmPFC. However, in the associated text, it says $p = .10$ so it is not significant. To avoid misleading readers, I would recommend pointing this out explicitly beyond saying 'most brain regions in the frontoparietal network also correlated with the intertemporal prior'.

Thank you for pointing this out. We now say on p.20

“With independent (leave-one-subject-out, LOSO) ROI analysis, we examined whether brain regions in the frontoparietal network (shown to represent strength of change evidence) correlated with intertemporal prior and found that all brain regions, with the exception of dmPFC, in the frontoparietal network correlated with the intertemporal prior.”

(4) There is a full paragraph in the discussion talking about the central opercular cortex, but this terminology has not shown up in the main body of the paper. If this is an important brain region to the authors, I would recommend mentioning it more often in the result section.

Thank you for this suggestion. We have now added central opercular cortex in the Results section (p.18):

“For P_r , we found that the ventromedial prefrontal cortex (vmPFC) and ventral striatum correlated with this behavioral measure of subjects’ belief about change. In addition, many

other brain regions, including the motor cortex, central opercular cortex, insula, occipital cortex, and the cerebellum also significantly correlated with P_t .”

(5) The authors have claimed that people make more extreme estimates under high diagnosticity (Supplementary Figure 1). This is an interesting point because it seems to be different from what is shown in the main graph where it seems that people are not extreme enough compared to an ideal Bayesian observer. I understand that these are effects being investigated under different circumstances. It would be helpful if for Supplementary Figure 1 the authors could overlay, or generate a different figure showing what an ideal Bayesian observer would do in this situation.

We thank the reviewer for pointing this out. We wish to clarify that when we said “more extreme estimates under high diagnosticity” we meant compared with low diagnosticity and not with the ideal Bayesian observer. We clarified this point by rephrasing our sentence on p.11:

“We also found that subjects tended to give more extreme P_t under high signal diagnosticity than low diagnosticity (Fig. S1 in Supplementary Information, SI).”

When it comes to comparing subjects’ probability estimates with the normative Bayesian, subjects tended to “underreact” under high diagnosticity. This can be seen in Fig. 4B, which shows a trend of increasing underreaction (or decreasing overreaction) as diagnosticity increased (row-wise comparison for a given transition probability).

We see the reviewer’s point in overlaying the Bayesian on Fig. S1 and update it by adding the normative Bayesian in orange.

<https://doi.org/10.7554/eLife.104684.2.sa0>