Criterion Setting is Implemented through Flexible Adjustment of Neural Excitability in Human Visual Cortex

Niels A. Kloosterman[*1,2,3,5], Jan Willem de Gee[2,4], Markus Werkle-Bergner[5], Ulman Lindenberger[1,5], Douglas D. Garrett[1,5+], Johannes Jacobus Fahrenfort[2,6+]

[1] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

[2] Department of Psychology, University of Amsterdam, The Netherlands;

[3] Center for Brain and Cognition, Institute for Interdisciplinary Studies, University of Amsterdam, The Netherlands;

[4] Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Germany;

[5] Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

[6] Department of Experimental and Applied Psychology, Vrije Universiteit, van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands

[+]Shared senior author

*Correspondence: kloosterman@mpib-berlin.mpg.de

Lead Contact Information

Niels A. Kloosterman, Ph.D.

Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Lentzeallee 94, 14195, Berlin, Germany

Phone: +49 30 82406 424

E-mail: kloosterman@mpib-berlin.mpg.de

Abstract

Choice bias, a hallmark of decision-making, is typically conceptualized as an internal reference, or “criterion”, against which accumulated evidence is compared. Flexible criterion adjustment allows organisms to adapt to the reward structure associated with the choice alternatives, and is assumed to arise from shifts in this reference. Here, in contrast, we show that criterion setting is implemented by modulating evidence accumulation rather than shifting an internal reference. Compared to a conservative criterion, experimentally inducing a liberal criterion during a visual detection task suppressed prestimulus oscillatory 8—12 Hz (alpha) activity in visual cortex, suggesting increased neural excitability. Increased excitability, in turn, boosted stimulus-related 59—100 Hz (gamma) activity by enhancing cortical response gain. Drift diffusion modeling of choice behaviour confirmed that a liberal criterion specifically biases the process of sensory evidence accumulation. Together, these findings provide a unique insight into the neural determinants of decision bias and its flexible adjustment.

Often, our decisions are not solely based on evaluation of available evidence. Instead, they can be heavily biased by pre-existing attitudes [1]. Such biases can be conceptualized as a flexible internal reference that dictates which decision to take, given the evidence. This reference, dubbed the criterion in signal detection theory (SDT) [2], is often set based on asymmetries in stimulus-response reward contingencies. Despite its critical role in decision making, the neural mechanism underlying criterion setting has yet to be identified. In principle, the criterion could be implemented as a flexible reference level of brain activity that is imposed on sensory representations, but the opposite scenario — in which not the reference but the

evidence accumulation process is flexibly regulated to implement criterion shifts — is equally plausible. Indeed, due to lack of a time dimension to track evidence accumulation, SDT cannot distinguish between these two possibilities [3,4]. Thus far, trial-to-trial variations in the criterion have been shown to correspond to spontaneous fluctuations in neural excitability, as measured in prestimulus oscillatory cortical activity in the 8—12 Hz (alpha) band [4-6]. Alpha oscillations, in turn, have been proposed to be involved in the gating of task-relevant sensory information [7], in line with adaptive regulation of evidence accumulation. No study, however, has directly demonstrated implementation of criterion setting through flexible adjustment of neural excitability and characterized the mechanism by which this is achieved.

Here, we asked whether experimentally induced criterion shifts during a visual target detection task result in corresponding changes in prestimulus alpha activity. Indeed, this approach showed that humans are able to intentionally alternate between conservative and liberal criteria by flexibly regulating neural excitability in visual cortex. Further, previous work has linked suppression of prestimulus alpha to heightened sensory responses by enhancing cortical response gain [8]. Accordingly, we found that increased neural excitability boosted high-frequency cortical responses to sensory stimulation. This suggests that, instead of changing a decision boundary, a liberal criterion increases the number of reported targets by biasing the process of sensory evidence accumulation towards this boundary. We finally tested this hypothesis using the drift diffusion model (DDM) [9], a sequential sampling model of decision making that can be viewed as an extension of SDT into the time domain [3]. This indeed confirmed that criterion setting specifically biases accumulation of sensory evidence, but not its starting point. Together, our findings show that criterion setting involves regulation of cortical excitability and evidence accumulation,

providing a unique insight into the neural underpinnings of decision bias and how it can flexibly be adjusted.

Results

Experimental task manipulations induce robust criterion shifts

Human participants viewed a continuous stream of full-screen textures (40 ms/texture) while their EEG was recorded. The orientation of the textures varied randomly, but the same orientation was never repeated twice in a row. The participants' task was to detect a central orientation-defined square, and report this target via a button press (Figure 1A). To ensure that targets and nontargets would appear in a consistent context, they appeared in a short fixed-order sequence within the continuous random stream. This fixed-order sequence (duration 1 s) always consisted of the same textures in the same order, with either a target or a nontarget appearing in the fifth position. The start of a fixed-order sequence within the continuous stream was unpredictable, but the onset of a recurring sequence of textures implicitly signalled an upcoming target or nontarget; the same fixed-order sequence reappeared throughout the experiment for every target and nontarget that was presented. In alternating nine-minute blocks, we manipulated the participants' criterion for reporting targets by instructing them either to report as many targets as possible ("Detect as many targets as possible!"; liberal condition), or to only report high-certainty targets ("Press only if you are really certain!"; conservative condition). Participants were free to respond at any time during a block whenever they detected a target. To support the task instructions, we provided auditory feedback following missed targets (misses) in the liberal condition and falsely detected targets (false

alarms) in the conservative condition, and applied monetary penalties for these errors (Figure 1A; see Methods for details).

Although the appearance of targets and nontargets was only implicitly signaled by the onset of a fixed-order sequence, reaction times (RT's) were clustered in time not only following target-present trials (hits), but also following target-absent trials (false alarms) (Figure S1A). This finding indicates that although trial onset was not explicitly cued, subjects in fact were sensitive to the onset of fixed-order sequences that signaled a potential target. The concurrent EEG recordings support this notion by revealing significant stimulus-related power modulations even in nontarget trials, in which a target was neither presented nor (falsely) detected (see next section and Figure S1B).

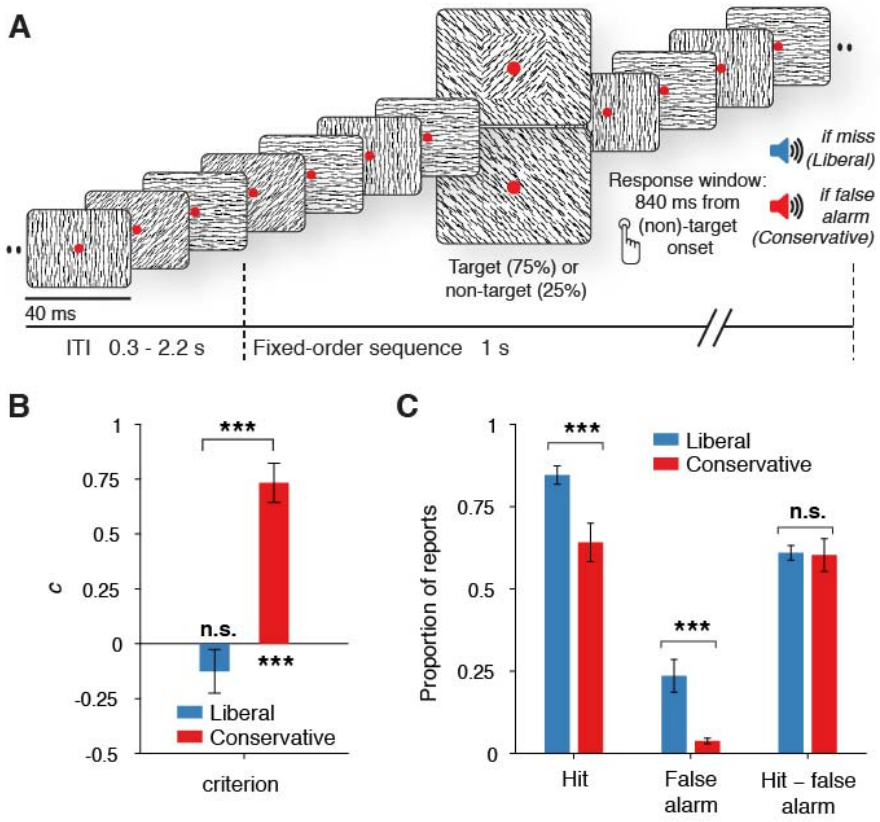


Figure 1 | Experimental task manipulations induce robust criterion shifts. A. Schematic of the visual stimulus and task design. Participants viewed an continuous stream of full-screen diagonally, horizontally and vertically oriented textures at a presentation rate of 40 ms (25 Hz). After random inter-trial intervals (range 0.3—2.2 s), a fixed-order sequence (duration 1 s) was presented, embedded in the stream. The fifth texture in each sequence either consisted of a single diagonal orientation (nontarget), or contained an orthogonal orientation-defined square (target of either 45° or 135° orientation). Participants decided whether they had just seen a target, reporting detected targets by button press within 840 ms after target onset. Liberal and conservative conditions were administered in alternating nine-minute blocks by penalizing either misses or false alarms, respectively, using aversive tones and monetary deductions. Depicted square and fixation dot sizes are not to scale. **B.** Signal-detection-theoretic (SDT) criterion during both conditions. **C.** Average proportion of hit and false alarm rates for each condition and average sensitivity quantified by the

difference between hit and false alarm rates. Error bars, SEM across participants (N = 16). ***p < 0.001; n.s., not significant.

Participants adjusted their criterion (c, quantified with SDT [2]; see Methods) depending on the experimental condition. The average criterion across participants was significantly greater than zero in the conservative condition (c = 0.73, all participants > 0, p < 0.0001, two-sided permutation test, 10,000 permutations), and slightly below zero in the liberal condition (c = -0.13, 11 out of 16 participants < 0, p = 0.22). Critically, c was significantly lower in the liberal than in the conservative condition for all participants (p < 0.0001). Given that c = 0 indicates a neutral criterion (i.e. no bias), this finding indicates that participants indeed adopted more liberal and more conservative criteria during both experimental conditions, while overall maintaining a relatively conservative attitude, in line with previous studies [10,11].

Compared to the subjective criterion, objective target detection performance was relatively similar in both conditions. Specifically, hit minus false alarm rates did not differ between the experimental conditions (p = 0.8185; Figure 1C). Observers were, however, less sensitive as quantified using SDT d' [2]; see Methods)) in the liberal condition: d' of 2.0 vs 2.31, p = 0.0002. Interestingly, reaction times were consistently shorter in the liberal condition too (mean RT liberal: 0.43 s (s.d. 0.03), conservative: 0.47 s (s.d. 0.05); p = 0.0002). Together, the decreased performance and faster responses in the liberal condition suggest that participants adapted their speed-accuracy tradeoff (SAT) depending on experimental condition to minimize losses, while detection ability remained similar [12]. Consistent with this idea, whereas c was on average 132% (s.d. 77) lower in the liberal than in the conservative

condition, *d'* decreased only 17% (s.d. 14)(*d'* vs. *c*: $p < 0.0001$), indicating that the experimental manipulations primarily affected the criterion. Taken together, the behavioral findings validate the experimentally induced criterion manipulation by demonstrating that participants adapted their criterion, thereby affecting their SAT while objective performance largely remained constant.

Task-relevant textures induce stimulus-related responses in visual cortex

Understanding the neural mechanisms underlying criterion setting first entails a careful examination of the stimulus-related signals on which the criterion is thought to operate. Such stimulus-related signals are typically reflected in visual cortical population activity exhibiting rhythmic temporal structure [13]. Specifically, bottom-up processing of visual information has previously been linked to increased high-frequency (> 40 Hz, i.e. gamma) electrophysiological activity over visual cortex [14-17]. Figure 2A shows time-frequency representations of EEG power modulations with respect to the prestimulus baseline period, recorded over visual cortex. Immediately following presentation of the task-relevant texture, we observed two distinct high-frequency power increases: one from 42—58 Hz reflecting the visual stimulation frequency (see below), and one from 59—100 Hz reflecting genuine gamma power modulation ($p < 0.05$, cluster-corrected for multiple comparisons, two-sided)(Figure 2A, top panel). The topography of this gamma modulation was confined to posterior electrodes (Figure 2B, top), in line with a role in visual cortical processing. Thus, the appearance of a task-relevant texture pattern in the continuous visual stream reliably induced a stimulus-related gamma response in visual cortex [18].

Concurrent with this gamma modulation, we observed several additional power modulations related to visual processing. First, spectral power increased in electrodes Pz and POz in a narrow frequency range around 25 Hz, reflecting the visual stimulation frequency of our experimental paradigm (Figure 2A, lower panel)[19]. This visual evoked potential was similarly expressed around 50 Hz – the first harmonic of the stimulation frequency (Figure 2A, top panel). Finally, we observed suppression of low-frequency (11—22 Hz) activity in posterior cortex, which typically occurs in parallel with enhanced stimulus-related gamma activity [20-23](Figure 2A, lower panel and 2B, right). Importantly, these low-frequency power modulations were significant even in correct rejection trials, in which a target was neither presented nor (falsely) detected (Figure S1B), indicating that participants were sensitive to the implicit structure of the task. To investigate criterion-related low-frequency power modulations in visual cortex before trial onset (described in the next section), we selected eleven posterior electrodes exhibiting enhanced stimulus-related gamma activity (highlighted electrodes in scalp maps in Figure 2B, top).

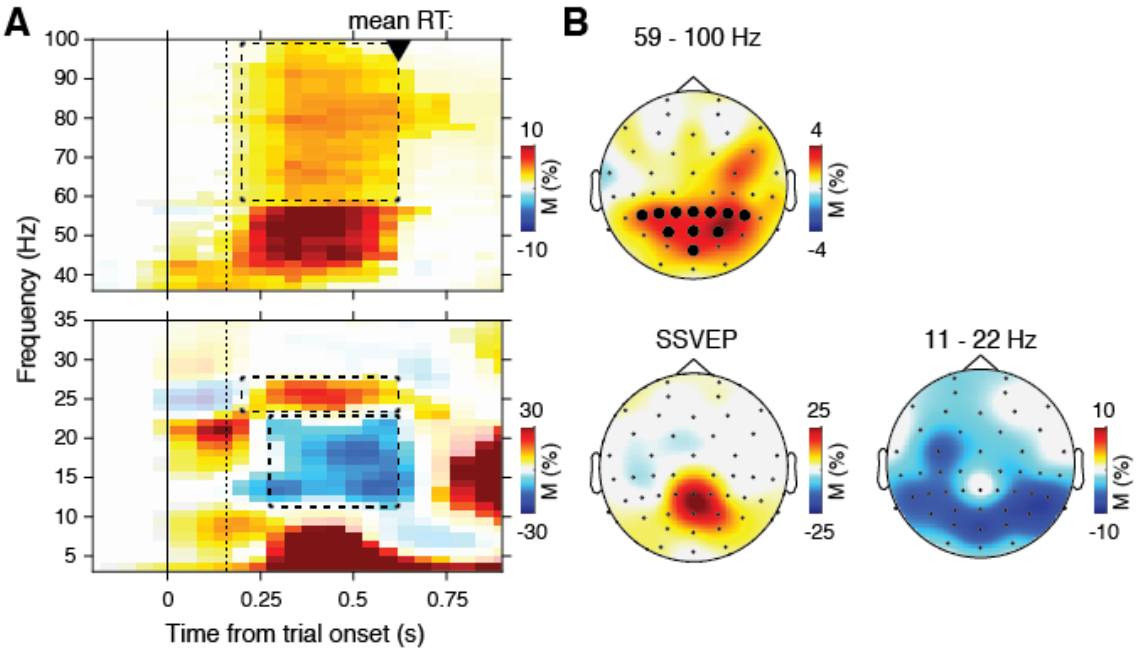


Figure 2 | Task-relevant textures induce stimulus-related responses in visual cortex. A. Time-frequency representations of high- (top) and low-frequency (bottom) EEG power modulations with respect to the prestimulus period (-0.4 - 0 s), pooled over the two conditions. Saturated colors indicate clusters of significant modulation, cluster threshold $p < 0.05$, two-sided permutation test across participants, cluster-corrected; N = 15). Solid and dotted vertical lines respectively indicate the onset of the trial and the target stimulus. M, power modulation. **B.** Scalp maps showing topography of stimulus-related modulations (0.2–0.6 s (gamma and SSVEP) or 0.25–0.6 s (low frequency suppression) after stimulus onset; see dashed outlines on time-frequency representations in A. Thick dots indicate electrodes used for the TFR's in A and which were selected for further analysis. SSVEP, steady state visual evoked potential.

Adopting a more liberal criterion increases neural excitability

Previous studies have shown that neural excitability, as reflected in suppression of prestimulus alpha (8—12 Hz) power, correlates with a decision-maker's propensity to categorize sensory input as a target [4,6]. This finding, however, does not elucidate the role of prestimulus alpha in criterion setting. Is the suppression of alpha activity merely a byproduct of a more liberal criterion, or do controlled criterion shifts change

neural excitability correspondingly, suggesting a critical role of alpha in criterion setting? To address this issue, we examined the effect of experimentally induced criterion shifts on alpha power between 0.8 and 0.2 s before trial onset.

The spatial topography of raw prestimulus alpha power pooled across the liberal and conservative conditions is plotted in Figure 3A. Alpha power was indeed strongest over our visual cortical electrode pooling of interest, suggesting involvement of this alpha activity in visual processing. Strikingly however, prestimulus alpha power was suppressed during the liberal compared to the conservative condition (Figure 3B), suggesting within-person modulation of alpha power during criterion setting. Expressing spectral power during the liberal condition as the percentage signal change from the conservative condition revealed a statistically significant cluster of suppressed frequencies precisely in the 8—12 Hz range ($p < 0.05$, cluster-corrected for multiple comparisons)(Figure 3D). This alpha suppression was located in a posterior region of the cerebral cortex largely overlapping with our visual electrode pooling of interest (Figure 3C). Taken together, these results show that experimentally induced criterion shifts are associated with a decrease of pre-stimulus alpha under a liberal criterion when compared to a conservative criterion. This finding suggests that alpha modulations are a hallmark of criterion setting, rather than having a spontaneously occuring haphazard influence on the criterion. Importantly, our finding suggests that humans are able to actively control the excitability of their brain in service of upcoming decisions.

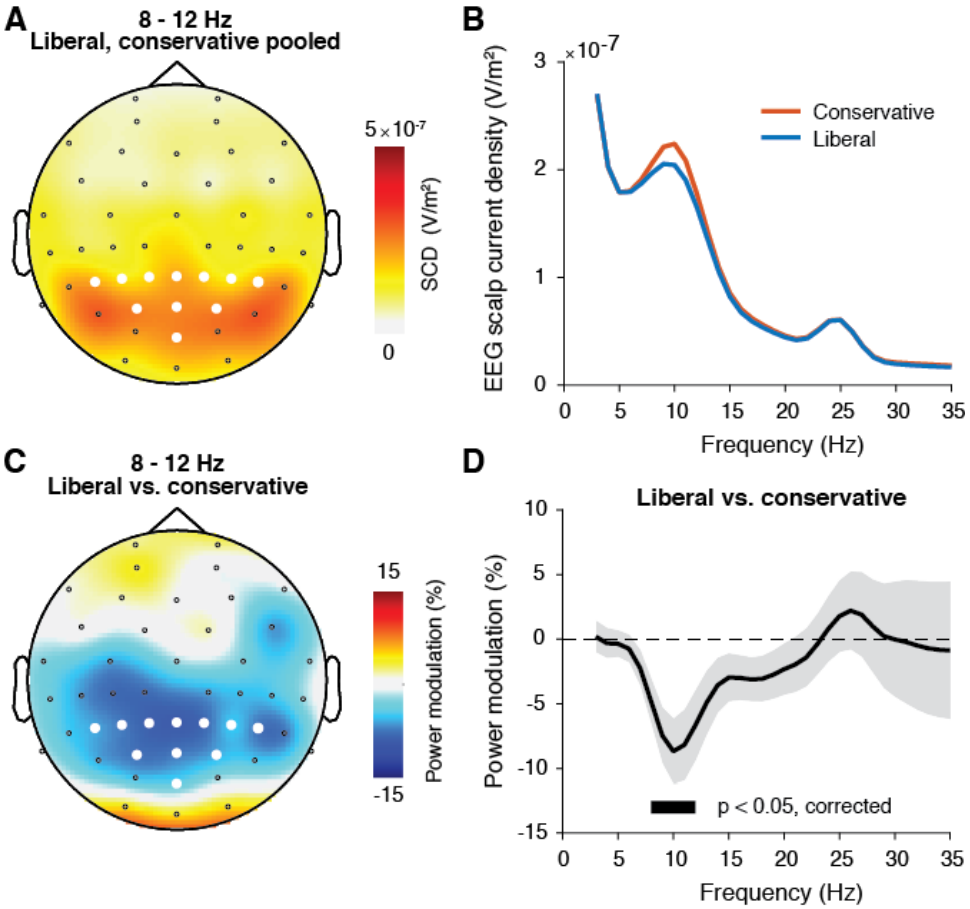


Figure 3 | Adopting a more liberal criterion increases neural excitability. A. Scalp map of raw prestimulus EEG alpha power (8—12 Hz neural activity between 0.8 and 0.2 s before sequence onset), pooled over conditions. White symbols indicate visual cortical electrodes used for the power spectra in B. and D. **B.** Low-frequency power spectra of prestimulus neural activity for both conditions. **C.** Scalp map of power modulation in the liberal condition, expressed as percent signal change from the conservative condition. **D.** Corresponding liberal versus conservative power spectrum. Black horizontal bar indicates statistically significant frequency range ($p < 0.05$, cluster-corrected for multiple comparisons, two-sided). Error bars, SEM across participants (N = 15).

Neural excitability boosts visual cortical responses by enhancing gain

How could increased neural excitability result in a more liberal criterion? One possibility is that increased excitability enhances stimulus-related activity in target as well as nontarget trials, thereby increasing the likelihood that this activity passes the

reference activity level for detecting a target. We explored this possibility using an existing theoretical framework that models the output activity of visual cortex as an s-shaped (sigmoidal) function of two factors: (1) the brain's excitability state in the current trial, here represented by prestimulus alpha power, and (2) stimulus-related neural activity, here represented by post-stimulus visual cortical gamma power [8] (Figure 4A). Assuming that stimulus-induced input activity is roughly similar across trials, the magnitude of the stimulus-related response in visual cortex is captured by the first derivative (i.e. the slope) of the sigmoidal function, yielding an inverted-U shaped response function (Figure 4B) [24,25]. Thus, according to the model, intermediate levels of neural excitability should produce the strongest stimulus-related activity. When heightened excitability in the liberal condition is observed, the model predicts enhanced stimulus processing (Figure 4B) in particular at intermediate prestimulus excitability levels through an increase in response gain [8] (i.e. steeper slope of the solid blue curve compared to the red curve in Figure 4B).

To test our prediction, we followed the method put forward by Rajagovindan and Ding [8]. We exploited the large number of trials per participant per condition in our study (range 543 to 1391 trials) by sorting each participant's trials per condition into ten excitability bins, as reflected in (log-transformed) prestimulus alpha power ranging from strong (indicating low excitability) to weak (indicating high excitability) (Figure 4C). We then averaged across the trials within each excitability bin the (log-transformed) stimulus-related visual cortical gamma power, and normalized each participant's binned gamma power by subtracting the lowest binned gamma observation in the conservative condition from all observations. Finally, we plotted the excitability bin number against the normalized gamma power, averaged across participants (see Methods for details).

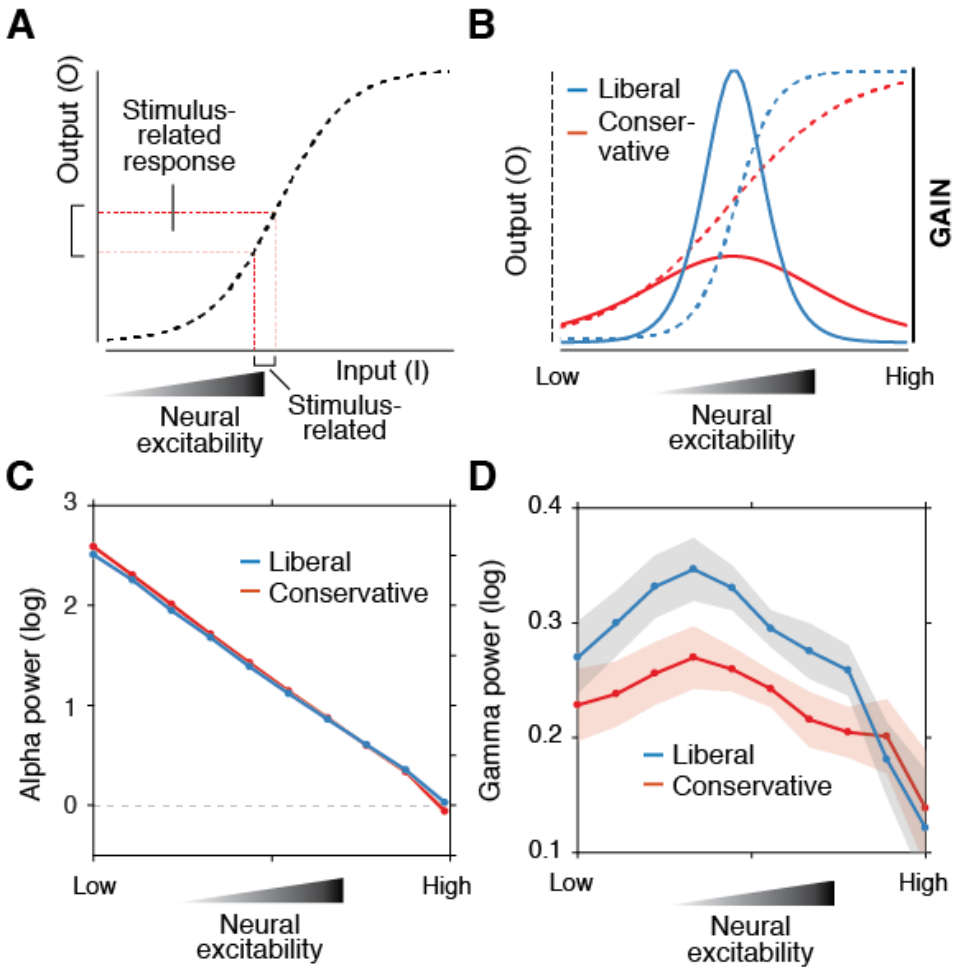


Figure 4 | Neural excitability boosts visual cortical responses by enhancing gain. A. Theoretical response gain model, which describes the transformation of stimulus-related input activity to output activity in visual cortex as a sigmoidal function, such that the current excitability state determines the output strength. **B.** Model predictions. Stimulus-related cortical output responses (solid lines) are formalized as the first derivative of the sigmoidal functions (dotted lines), resulting in an inverse-U shaped response gain function. The model predicts that a liberal criterion increases the steepness of the sigmoidal function (right) compared to a conservative criterion (left), resulting in stronger stimulus-related responses [8]. **C.** Neural excitability reflected in single trial, log-transformed prestimulus alpha power sorted across ten bins from high (indicating low excitability) to low (indicating high excitability), separately for both conditions. **D.** Corresponding log-transformed gamma activity (normalized within participants by subtracting the minimum gamma power during the conservative condition from all

bins) plotted as a function of neural excitability. Error bars, within-subject SEM across participants (N = 14).

The resulting plot closely resembles an inverted-U shaped relationship between excitability and stimulus-related gamma activity for both conditions, with particularly low gamma responses for high excitability trials (Figure 4D). Critically, average gamma activity was higher for the liberal than for the conservative condition, except during the highest excitability states (Figure 4D, rightmost two data points). Indeed, the flanks of the inverted-U curve for the liberal condition were steeper for the liberal condition, suggesting increased response gain. A three-way repeated measures ANOVA with factors condition (conservative, liberal), brain activity type (prestimulus alpha, poststimulus gamma) and bin level (1-10) revealed a significant three-way interaction ($F(9,117) = 2.96$, $p = 0.003$, partial $\eta^2 = 0.19$, Greenhouse-Geisser corrected $p = 0.046$). Importantly, the marginally significant quadratic contrast ($F(1,13) = 3.47$, $p = 0.085$, partial $\eta^2 = 0.21$) fitted this interaction almost as well as a linear contrast ($F(1,13) = 4.69$, $p = 0.049$, partial $\eta^2 = 0.265$). This three-way quadratic interaction effect indeed suggests a more steeply U-shaped curve for gamma responses in the liberal condition, in line with enhanced gain. Taken together, these findings suggest that the increased excitability during the liberal condition boosted sensory stimulus processing. In turn, this boosted activity might have increased the participant's propensity to categorize both target and nontarget stimuli as signal, resulting in a more liberal response attitude.

Evidence accumulation bias underlies experimentally induced liberal criterion

Finally, we used computational modeling of our behavioral data to investigate whether an experimentally induced liberal criterion specifically affects the process of sensory evidence accumulation, as suggested by our observed enhancement of stimulus-related activity. To this end, we fitted our reaction time data (Figure S1A) using the drift diffusion model (DDM), an established dynamic model of two-choice decision processes [9](Figure 5A). The DDM postulates that decisions are reached by accumulation of noisy sensory evidence towards one of two decision boundaries, which can either be explicit (for "yes" responses in our experiment), or implicit (i.e. without active response, for "no" decisions in our experiment) [26,27] . Within this model, a liberal decision bias can emerge in two different ways: either by moving the starting point of evidence accumulation closer to the "yes" decision boundary ("starting point" parameter), or by driving the evidence accumulation process itself more towards the "yes" boundary ("drift criterion" parameter, equivalent to a bias in evidence accumulation, implemented by adding an evidence-independent constant to the drift). Previous work has shown that an evidence accumulation bias drives a reduced conservative bias (i.e. a more liberal attitude) during pupil-linked arousal [10].

To test whether an evidence accumulation bias similarly underlies decision bias during an experimentally induced liberal criterion, we fitted the model while freeing various parameters per experimental condition (drift rate, boundary separation, non-decision time and drift criterion) and keeping starting point fixed. This model ("drift criterion model") fitted our single-participant data well (Figures 5B and S2). Indeed, actual drift rate was significantly lower in the liberal than in the conservative condition ($p < 0.0001$, permutation test), supported by a highly positive drift criterion parameter during the liberal condition and a null-level parameter for the conservative condition ($p = 0.90$ for conservative; liberal vs, conservative: $p <$

0.0001; Figure 5C, top). Strikingly, the condition-induced shifts in SDT criterion and DDM drift criterion were strongly correlated across participants (Pearson's r = – 0.89, n = 16, p = $4.1e^{-6}$), indicating that these two measures reflect similar aspects of our data (Figure 5D). In addition, starting point (fixed across conditions) was below 0.5, indicating that this parameter was closer to the "no" boundary overall, in line with a generally conservative response attitude ($p < 0.0001$), as also observed in SDT criterion (compare Figures 1B and 5C, bottom left). Finally, boundary separation also increased slightly (but reliably) during the liberal condition (p = 0.0001) (Figure 5C, top left), whereas non-decision time decreased (p = 0.0001) (Figure 5C, bottom right).

As a first control of the goodness of fit of the condition-dependent effect on drift criterion, we re-fitted the model while fixing both drift criterion and starting point for both conditions, but still allowing all other of the above (non-bias-related) parameters to vary freely. This "basic model" provided a worse fit to the data, as indicated by higher Bayesian Information Criterion (BIC) estimates than for the drift criterion model (Figure 5E)(see Methods for details). As a second control, we fitted the model again while fixing drift criterion for both experimental conditions, and instead allowing starting point and all other of the above parameters to vary ("starting point model"). This model also provided a worse fit to the data (Figure 5E). Specifically, for 12 out of 16 participants the drift criterion model provided better fits to behavior than the starting point model, for 10 of which delta BIC was greater than 10 (indicating very strong evidence against the starting point model). Taken together, our modelling results suggest that participants achieved a liberal decision bias specifically by biasing the evidence accumulation process towards "yes" decisions (not its starting point), while non-bias-related parameters were also affected.

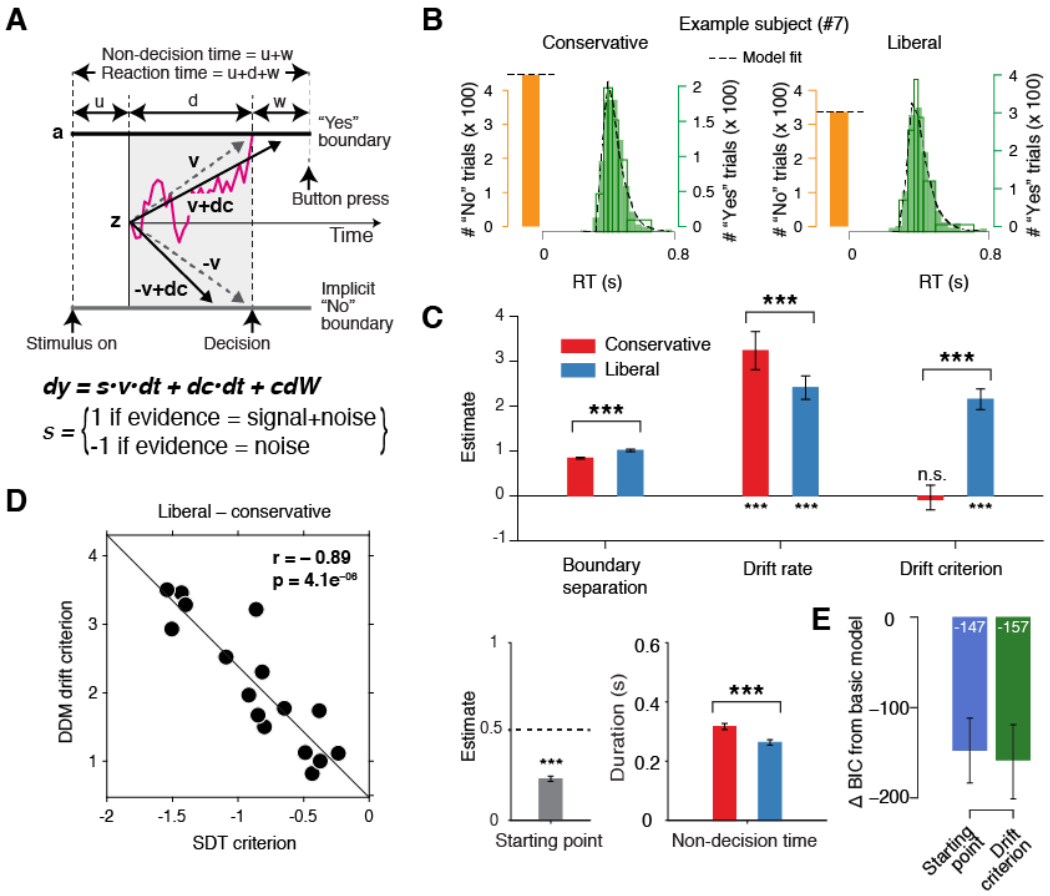


Figure 5 | Evidence accumulation bias underlies experimentally induced liberal criterion. A. Schematic and simplified equation of drift diffusion model accounting for RT distributions for explicit "yes"- and implicit "no"-choices ('stimulus coding'; see Methods). Notation: dy, change in decision variable y per unit time dt; v·dt, mean drift (multiplied with 1 for signal+noise (target) trials, and -1 for noise (non-target) trials); dc·dt, drift criterion (an evidence-independent constant added to the drift); and cdW, Gaussian white noise (mean = 0, variance = c2·dt). **B.** RT distributions of an example subject for "yes"- choices and the number of implicit "no" choices, separately for the two conditions. See Figure S2 for all participant data. Green bars, observed RT quantiles. **C.** Estimated model parameters for the drift criterion model. Conventions as in Figure 1. N = 16 participants. **D.** Across-participant Pearson correlation between experimentally induced (Liberal—conservative) shifts in SDT criterion and DDM drift criterion. **E.** BIC goodness of fit estimates for the starting point and drift criterion models, expressed with respect to a basic model without bias parameters. A lower delta BIC value indicates a better fit.

Discussion

Our subjective criterion for choosing a course of action plays a central role in every decision we make. To date, however, the neural underpinnings of criterion setting have remained elusive. Here, we demonstrated that instructed modulations of an observer's criterion resulted in robust changes in the excitability state of the brain, as reflected in prestimulus alpha activity over posterior cortex. Revealing the underlying mechanism, we showed that a more excitable prestimulus brain state associated with a liberal criterion boosted stimulus-related high-frequency cortical responses over visual cortex by increasing response gain. Drift diffusion modeling of our behavioral data confirmed that criterion shifts are achieved by biasing the evidence accumulation process. Together, these findings show that humans implement decision biases by flexibly adapting neural excitability and sensory evidence accumulation.

One neural mechanism that could underlie this enhanced processing may be under control of the catecholaminergic neuromodulatory systems, consisting of the noradrenaline-releasing locus coereleus (LC) and dopamine systems [28]. These systems are able to modulate the level of arousal and neural gain, and show tight links with pupil responses [10,11,29,30]. Accordingly, prestimulus alpha power suppression has also recently been linked to pupil dilation [21]. From this perspective, our results reconcile previous studies showing relationships between a liberal criterion, suppression of spontaneous alpha power and increased pupil size. Consistent with this, a recent monkey study observed increased neural activity during a liberal criterion in the superior colliculus [31], a mid-brain structure interconnected with the LC [29]. Taken together, a more liberal within-person criterion (following experimental instruction) might activate neuromodulatory systems that

subsequently increase cortical excitability and enhance sensory responses for both stimulus and “noise” signals in visual cortex, thereby increasing a person’s propensity for “yes” responses [4].

One difference relative to previous studies investigating the link between neural excitability and decision criterion is that previous experiments contained active “no” decisions [4,6], whereas our participants only executed active “yes” decisions (“no” decisions were implicit). This design was chosen to create a naturalistic setting in which participants were encouraged to continuously detect targets. Although our finding of increased neural excitability during a liberal criterion is consistent with previous studies, future studies could include active “no” decisions to test whether a liberal response attitude increases excitability similarly when both choice alternatives entail active responses. In any case, despite the naturalistic form of our continuous experimental paradigm, both hits and false alarms showed typically observed reaction time distributions (Figure S1A), suggesting that participants were well able to extract the implicit task structure. The tight drift diffusion model fits (Figure S2), as well as the low-frequency EEG power modulations (Figure S1B) further confirm this notion. Regarding our observed shorter RT’s in the liberal condition, we speculate that a more liberal response attitude might in general be associated with an emphasis on speed versus accuracy. As previous criterion studies do not report reaction times, future studies should further address this issue by taking into account the temporal aspect of evidence accumulation, as is done in drift diffusion models.

Although one could argue that observed change in cortical excitability may reflect a change in detection sensitivity rather than an intentional criterion shift, we deem this scenario unlikely because it invokes effects opposite to those we

observed. We found increased excitability in the liberal condition when compared to the conservative condition. If this were due to improved detection performance, one would predict higher sensitivity in the liberal condition, while in fact we found higher sensitivity in the conservative condition. This finding convincingly ties cortical excitability in our paradigm to a strategically applied criterion shift, as opposed to a change in detection sensitivity. Convergently, other studies also report a link between prestimulus low-frequency EEG activity and subjective perception, but not objective task performance [32,33].

Finally, we do not believe that the observed high-frequency EEG power modulation in the current study was driven by microsaccade-related activity [34] for several reasons. First, our highly dynamic and salient visual stimulus was optimal to elicit high-frequency responses, as shown with MEG, a technique less susceptible to microsaccades [35,36]. Second, we reduced the sensitivity of the EEG to microsaccades by using the earlobes instead of the nose as the reference [34], by removing microsaccade-related activity from the data [37,38], and by applying a current source density transformation to our data [39](see Methods for details). Third, the onset of a task-relevant event is often associated with a drop in microsaccade rate [40,41], which should have resulted in a corresponding drop in high-frequency power; conversely, we find that high frequency power increases at target onset. Future work could further address this issue by measuring eye movements during the experimental tasks.

Overall, our results suggest that stimulus-related responses are boosted during a liberal criterion due to increased cortical response gain, which is further supported by recent work linking alpha power suppression to enhanced gain [42]. Explicit manipulation of cortical response gain during a criterion manipulation by

pharmacological manipulation of the noradrenergic LC-NE system [43] or by enhancing occipital alpha power using transcranial stimulation [44] would further establish the underlying mechanisms involved in criterion setting. In the end, although one may be unaware, every decision we make is influenced by implicit biases that operate on the noisy evidence accumulation process towards one of the alternatives. Understanding how these biases affect our decisions is key to becoming aware of these biases [45], allowing us to control or invoke them adaptively. We argue that pinpointing the neural mechanisms underlying bias in an elementary perceptual task (as used here) paves the way for understanding how more abstract and high-level decisions may be modulated by choice bias [46].

Acknowledgments

This work was supported by the Max Planck Society. In addition, DDG and NAK are supported by an Emmy Nöther Programme grant from the German Research Foundation (to DDG) and by the Max Planck UCL Centre for Computational Psychiatry and Ageing Research. MWB is supported by a grant from the German Research Foundation (DFG; WE4296/5-1), as well as the Jacobs Foundation via an Early Career Research Fellowship.

Author Contributions

NAK and JJF designed research, NAK performed research, NAK, JWdG and JJF analyzed data, MWB and DDG provided theoretical background, NAK, JJF, UL, MWB, DDG, JWdG wrote the paper, NAK, JJF, UL, MWB, DDG, JWdG edited and commented on the manuscript.

Declaration of Interests

The authors declare no competing interests.

References

1. Gross, M. Can we change our biased minds? *Curr Biol* **27,** R1089–R1091 (2017).
2. Green, D. M. & Swets, J. A. Signal detection theory and psychophysics. *Society* **1,** 521 (1966).
3. Fetsch, C. R., Kiani, R. & Shadlen, M. N. Predicting the Accuracy of a Decision: A Neural Mechanism of Confidence. *Cold Spring Harb. Symp. Quant. Biol.* **79,** 185–197 (2014).
4. Iemi, L., Chaumon, M., Crouzet, S. M. & Busch, N. A. Spontaneous Neural Oscillations Bias Perception by Modulating Baseline Excitability. *J Neurosci* **37,** 807–819 (2017).
5. Samaha, J., Iemi, L. & Postle, B. R. Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy. *Consciousness and Cognition* (2017). doi:10.1016/j.concog.2017.02.005
6. Limbach, K. & Corballis, P. M. Prestimulus alpha power influences response criterion in a detection task. *Psychophysiology* **53,** 1154–1164 (2016).
7. Jensen, O. & Mazaheri, A. Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Front Hum Neurosci* **4,** 186 (2010).
8. Rajagovindan, R. & Ding, M. From prestimulus alpha oscillation to visual-evoked response: an inverted-U function and its attentional modulation. *J Cogn Neurosci* **23,** 1379–1394 (2011).
9. Ratcliff, R. & McKoon, G. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Comput* **20,** 873–922 (2008).
10. de Gee, J. W. *et al.* Dynamic modulation of decision biases by brainstem arousal systems. *eLife Sciences* **6,** 309 (2017).
11. de Gee, J. W., Knapen, T. & Donner, T. H. Decision-related pupil dilation reflects upcoming choice and individual bias. *Proc. Natl. Acad. Sci. U.S.A.* **111,** E618–25 (2014).
12. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev* **113,** 700–765 (2006).
13. Buzsáki, G. & Draguhn, A. Neuronal oscillations in cortical networks. *Science* **304,** 1926–1929 (2004).
14. Popov, T., Kastner, S. & Jensen, O. FEF-Controlled Alpha Delay Activity Precedes Stimulus-Induced Gamma-Band Activity in Visual Cortex. *Journal of Neuroscience* **37,** 4117–4127 (2017).

15. Bastos, A. M. *et al.* Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron* **85,** 390–401 (2015).
16. van Kerkoerle, T. *et al.* Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 14332–14341 (2014).
17. Michalareas, G. *et al.* Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron* **89,** 384–397 (2016).
18. Ni, J. *et al.* Gamma-Rhythmic Gain Modulation. *Neuron* **92,** 240–251 (2016).
19. Müller, M. M., Teder, W. & Hillyard, S. A. Magnetoencephalographic recording of steadystate visual evoked cortical activity. *Brain Topogr* **9,** 163–168 (1997).
20. Kloosterman, N. A. *et al.* Top-down modulation in human visual cortex predicts the stability of a perceptual illusion. *J Neurophysiol* **113,** 1063–1076 (2015).
21. Meindertsma, T., Kloosterman, N. A., Nolte, G., Engel, A. K. & Donner, T. H. Multiple Transient Signals in Human Visual Cortex Associated with an Elementary Decision. *Journal of Neuroscience* **37,** 5744–5757 (2017).
22. Donner, T. H. & Siegel, M. A framework for local cortical oscillation patterns. *Trends Cogn Sci* **15,** 191–199 (2011).
23. Werkle-Bergner, M. *et al.* Coordinated within-Trial Dynamics of Low-Frequency Neural Rhythms Controls Evidence Accumulation. *Journal of Neuroscience* **34,** 8519–8528 (2014).
24. Destexhe, A., Rudolph, M., Fellous, J. M. & Sejnowski, T. J. Fluctuating synaptic conductances recreate in vivo-like activity in neocortical neurons. *Neuroscience* **107,** 13–24 (2001).
25. Freeman, W. J. Nonlinear gain mediating cortical stimulus-response relations. *Biol Cybern* **33,** 237–247 (1979).
26. de Gee, J. W., Tsetsos, K., McCormick, D. A., McGinley, M. J. & Donner, T. H. Phasic pupil-linked arousal reduces decision biases in mice and men. *Soc for Neurosci abstr* **713.07 / UU49,** (2017).
27. Ratcliff, R., Huang-Pollock, C. & McKoon, G. Modeling Individual Differences in the Go/No-Go Task With a Diffusion Model. *psycnet.apa.org* (2016). doi:http:// dx.doi.org/10.1037/dec0000065
28. Aston-Jones, G. & Cohen, J. D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu Rev Neurosci* **28,** 403–450 (2005).
29. Joshi, S., Li, Y., Kalwani, R. M. & Gold, J. I. Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron* **0,** 221–234 (2015).
30. McGinley, M. J., David, S. V. & McCormick, D. A. Cortical Membrane Potential Signature of Optimal States for Sensory Signal Detection. *Neuron* **87,** 179–192 (2015).
31. Crapse, T. B., Lau, H. & Basso, M. A. A Role for the Superior Colliculus in Decision Criteria. *Neuron* **97,** 181–194.e6 (2018).
32. Iemi, L. & Busch, N. A. Moment-to-moment fluctuations in neuronal excitability bias subjective perception rather than decision-making. *bioRxiv* 151324 (2017). doi:10.1101/151324

33. Benwell, C. S. Y. *et al.* Pre-stimulus EEG power predicts conscious awareness but not objective visual performance. *eNeuro* **4,** ENEURO.0182–17.2017 (2017).
34. Yuval-Greenberg, S., Tomer, O., Keren, A. S., Nelken, I. & Deouell, L. Y. Transient Induced Gamma-Band Response in EEG as a Manifestation of Miniature Saccades. *Neuron* **58,** 429–441 (2008).
35. Siegel, M., Donner, T. H., Oostenveld, R., Fries, P. & Engel, A. K. High-Frequency Activity in Human Visual Cortex Is Modulated by Visual Motion Strength. *Cerebral Cortex* **17,** 732–741 (2006).
36. Hoogenboom, N., Schoffelen, J.-M., Oostenveld, R., Parkes, L. M. & Fries, P. Localizing human visual gamma-band activity in frequency, time and space. *Neuroimage* **29,** 764–773 (2006).
37. Hassler, U., Trujillo-Barreto, N. & Gruber, T. Induced gamma band responses in human EEG after the control of miniature saccadic artifacts. *Neuroimage* **57,** 1411–1421 (2011).
38. Hipp, J. F. & Siegel, M. Dissociating neuronal gamma-band activity from cranial and ocular muscle activity in EEG. *Front Hum Neurosci* **7,** 338 (2013).
39. Melloni, L., Schwiedrzik, C. M., Wibral, M., Rodriguez, E. & Singer, W. Response to: Yuval-Greenberg et al., 'Transient Induced Gamma-Band Response in EEG as a Manifestation of Miniature Saccades.' Neuron 58, 429-441. *Neuron* **62,** 8–10– author reply 10–12 (2009).
40. Bonneh, Y. S. *et al.* Motion-induced blindness and microsaccades: Cause and effect. *J Vis* **10,** 22–22 (2010).
41. Rolfs, M. Microsaccades: small steps on a long way. *Vis. Res.* **49,** 2415–2441 (2009).
42. Peterson, E. J. & Voytek, B. Alpha oscillations control cortical gain by modulating excitatory-inhibitory background activity. *biorxiv.org* (2017). doi:https://doi.org/10.1101/185074
43. Servan-Schreiber, D., Printz, H. & Cohen, J. D. A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science* **249,** 892–895 (1990).
44. Zaehle, T., Rach, S. & Herrmann, C. S. Transcranial Alternating Current Stimulation Enhances Individual Alpha Activity in Human EEG. *PLoS ONE* **5,** e13766 (2010).
45. Pleskac, T. J., Cesario, J. & Johnson, D. J. How race affects evidence accumulation during the decision to shoot. *Psychon Bull Rev* **18,** 1–30 (2017).
46. Tversky, A. & Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science* **185,** 1124–1131 (1974).
47. Fahrenfort, J. J., Scholte, H. S. & Lamme, V. A. F. Masking disrupts reentrant processing in human visual cortex. *J Cogn Neurosci* **19,** 1488–1497 (2007).
48. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J.-M. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience* **2011,** 1–9 (2011).
49. Perrin, F., Pernier, J., Bertrand, O. & Echallier, J. F. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology* **72,** 184–187 (1989).

50. Mitra, P. P. & Pesaran, B. Analysis of Dynamic Brain Imaging Data. *Biophysical Journal* **76,** 691–708 (1999).
51. Wiecki, T. V., Sofer, I. & Frank, M. J. HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front. Neuroinform.* **7,** (2013).
52. Neath, A. A. & Cavanaugh, J. E. The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics* **4,** 199–203 (2012).
53. Efron, B. & Tibshirani, R. The problem of regions. *The Annals of Statistics* **26,** 1687–1718 (1998).
54. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Methods* **164,** 177–190 (2007).

Materials and Methods

Participants Sixteen participants (eight female, mean age 24.1 years, ± 1.64) took part in the experiment, either for financial compensation or in partial fulfillment of first year course requirements. Each participant completed three experimental sessions on different days, each session lasting ca. 1.5 hours, including preparation and breaks. One participant completed only two sessions, yielding a total number of one-hour measurements across subjects of 47. Due to technical issues, for one session only data for the liberal condition was available. One participant was an author. All participants were included in the behavioral and drift diffusion modeling analyses. One participant was excluded from the stimulus-related (Figure 2) and the alpha-power analysis (Figure 3) due to excessive noise (EEG power spectrum opposite of 1/frequency). One further participant was excluded from the single-trial gamma power modulation analyses (Figure 4) because the liberal-conservative difference in gamma power in this participant was > 3 standard deviations away from the other participants. In summary, 16 participants were included in the analyses presented in Figures 1, 5 and S2, 15 participants in Figures 2, 3 and S2, and 14 participants in Figure 4. All participants had normal or corrected-to-normal vision and were right handed. Participants provided written informed consent before the start of the

experiment. All procedures were approved by the ethics committee of the University of Amsterdam.

Stimuli Stimuli consisted of a continuous semi-random rapid serial visual presentation (rsvp) of full screen texture patterns. The texture patterns consisted of line elements approx. 0.07° thick and 0.4° long in visual angle. Each texture in the rsvp was presented for 40 ms (i.e. stimulation frequency 25 Hz), and was oriented in one of four possible directions: 0°, 45°, 90° or 135°. Participants were asked to fixate on a red dot in the center of the screen. At random inter trial intervals (ITI's) sampled from a uniform distribution (ITI range 0.3-2.2 s), the rsvp contained a fixed sequence of 25 texture patterns, which in total lasted one second. This fixed sequence consisted of four stimuli preceding a (non-)target stimulus (orientations of 45°, 90°, 0°, 90° respectively) and twenty stimuli following the (non)-target (orientations of 0°, 90°, 0°, 90°, 0°, 45°, 0°, 135°, 90°, 45°, 0°, 135°, 0°, 45°, 90°, 45°, 90°, 135°, 0°, 135° respectively) (Figure 1). The fifth texture pattern within the sequence (occurring from 0.16 s after sequence onset) was either a target or a nontarget stimulus. Nontargets consisted of either a 45° or a 135° homogenous texture, whereas targets contained a central orientation-defined square of 2.42° visual angle, thereby consisting of both a 45° and a 135° texture. 50% of all targets consisted of a 45° square and 50% of a 135° square. Of all trials, 75% contained a target and 25% a nontarget. Target and nontarget trials were presented in random order. To avoid specific influences on target stimulus visibility due to presentation of similarly or orthogonally oriented texture patterns temporally close in the cascade, no 45° and 135° oriented stimuli were presented directly before or after presentation of the target stimulus. All stimuli had an isoluminance of 72.2 cd/m^2. Stimuli were created

using MATLAB (The Mathworks, Inc., Natick, MA, USA) and presented using Presentation (Neurobehavioral systems, Inc., Albany, CA, USA).

Experimental design The participants' task was to detect targets and actively report them by pressing a button using their preferred hand. Presumably due to constant forward and backward masking by the continuous cascade of stimuli and unpredictability of target timing, targets occasionally went unreported [47]. The onset of the fixed order of texture patterns preceding and following (non-)target stimuli was neither signaled nor apparent.

At the beginning of the experiment, participants were informed they could earn a total bonus of EUR 30, on top of their regular pay or course credit. In two separate conditions within each session of testing, we encouraged participants to use either a conservative or a liberal criterion for reporting targets using both aversive sounds as well as reducing their bonus after errors. In the conservative condition, participants were instructed to only press the button when they were relatively sure they had seen the target. The instruction on screen before block onset read as follows: "Try to detect as many targets as possible. Only press when you are relatively sure you just saw a target." To maximize effectiveness of this instruction, participants were told the bonus would be diminished by ten cents after a false alarm. During the experiment, a loud aversive sound was played after a false alarm to inform the participant about an error. During the liberal condition, participants were instructed to miss as few targets as possible. The instruction on screen before block onset read as follows: "Try to detect as many targets as possible. If you sometimes press when there was nothing this is not so bad". In this condition, the loud aversive sound was played twice in close succession whenever they failed to report a target, and three cents were subsequently deducted from their bonus. The difference in

auditory feedback between both conditions was included to inform the participant about the type of error (miss or false alarm), in order to facilitate the desired criteria in both conditions. After every block, the participant's score (number of missed targets in the liberal condition and number of false alarms in the conservative condition) was displayed on the screen, as well as the remainder of the bonus. After completing the last session of the experiment, every participant was paid the full bonus as required by the ethical committee.

During a block, participants continuously monitored the screen and were free to respond by button press whenever they thought they saw a target. Participants performed six blocks per session. Each block contained 240 trials, 180 target and 60 nontarget trials. The task instruction was presented on the screen before the block started. The condition of the first block of a session was counterbalanced across participants. Prior to EEG recording in the first session, participants performed a 10-minute practice run of both conditions, in which visual feedback directly after a miss (liberal condition) or false alarm (conservative) informed participants about their mistake, allowing them to adjust their decision criterion accordingly.

Behavioral analysis We calculated participants criterion *c* [2] across the trials in each condition as follows:

$$c = -\frac{1}{2}\left[Z(Hit\text{-}rate) + Z(FA\text{-}rate)\right]$$

where Z(...) is the inverse standard normal distribution. Furthermore, we calculated objective sensitivity measure d' using:

$$d' = Z(Hit\text{-}rate) - Z(FA\text{-}rate)$$

as well as by subtracting hit and false alarm rates. Reaction times (RT's) were measured as the period between target onset and button press.

EEG recording Continuous EEG data were recorded at 256 Hz using a 48-channel BioSemi Active-Two system (Biosemi, the Netherlands), connected to a standard EEG cap according to the international 10-20 system. Electrooculography (EOG) was recorded using two electrodes at the outer canthi of the left and right eyes and two electrodes placed above and below the right eye. Horizontal and vertical EOG electrodes were referenced against each other, two for horizontal and two for vertical eye movements (blinks). We used the Fieldtrip toolbox [48] and custom software in MATLAB (version R2016b, The Mathworks) to process the data (see below). Data were re-referenced to the average voltage of two electrodes attached to the earlobes.

Trial extraction and preprocessing We extracted trials of variable duration from 1 s before target sequence onset until 1.25 after button press for trials that included a button press (hits and false alarms), and until 1.25 s after stimulus onset for trials without a button press (misses and correct rejects). The following constraints were used to classify (non-)targets as detected (hits and false alarms), while avoiding the occurrence of button presses in close succession to target reports and button presses occurring outside of trials: 1) A trial was marked as detected if a response occurred within 0.8 s after target offset; 2) when the onset of the the next target stimulus sequence started before trial end, the trial was terminated at next trials onset; 3) when a button press occurred in the 1.5 s before trial onset, the trial was extracted from 1.5 s after this button press; 4) when a button press occurred between 0.5 s before until 0.2 s after sequence onset, the trial was discarded. See

Kloosterman et al. [20] and Meindertsma et al. [21] for a similar trial extraction procedure. After trial extraction, channel time courses were linearly detrended and the mean of every channel was removed per trial.

Artifact rejection Trials containing muscle artifacts were rejected from further analysis using a standard semi-automatic preprocessing method in Fieldtrip. This procedure consists of bandpass-filtering the trials of a condition block in the 110–125 Hz frequency range, which typically contains most of the muscle artifact activity, followed by a Z-transformation. Trials exceeding a threshold Z-score were removed completely from analysis. We used as the threshold the absolute value of the minimum Z-score within the block, + 1. To remove eye blink artifacts from the time courses, the EEG data from a complete session were transformed using independent component analysis (ICA), and components (typically one or two of the 48) due to blinks was removed from the data. In addition, to remove microsaccade-related artifacts we included two virtual channels based on channels Fp1 and Fp2 in the ICA, which included transient spike potentials as identified using the algorithm from Hassler et al. [37]. The two components loading high on these virtual electrodes were also removed. Blinks and eye movements were then semi-automatically detected from the horizontal and vertical EOG (frequency range 1–15 Hz; z-value cut-off 4 for vertical; 6 for horizontal) and trials containing eye artefacts within 0.1 s around target onset were discarded. This step was done to remove trials in which the target was not seen because the eyes were closed. Finally, trials exceeding a threshold voltage range of 200 μV were discarded. To attenuate volume conduction effects and suppress any remaining microsaccade-related activity, the scalp current density (SCD) was computed using the second-order derivative (the surface Laplacian) of the EEG potential distribution [49].

Spectral analysis of EEG power We used a sliding window Fourier transform ([50]; step size, 50 ms; window length, 400 ms; frequency resolution, 2.5 Hz) to calculate time-frequency representations (spectrograms) of the EEG power for each electrode and each trial. We used a single Hann taper for the frequency range of 3–35 Hz (spectral smoothing, 4.5 Hz, bin size, 1 Hz) and the multitaper technique for the frequency range of 36 – 100 Hz (spectral smoothing, 8 Hz; bin size, 2 Hz; five tapers). See Kloosterman et al. [20] and Meindertsma et al. [21] for similar settings.

Spectrograms were aligned to the onset of the stimulus sequence containing the (non)target. Power modulations (denoted as M in Figure 2) during the trials were quantified as the percentage of power change at a given time point and frequency bin, relative to a baseline power value for each frequency bin. We used as a baseline the mean EEG power in the interval 0.4 to 0 s before trial onset. If this interval was not completely present in the trial due to preceding events (see Trial extraction), this period was shortened accordingly. We subtracted the trial-specific baseline value from each sample in the time course per frequency bin and divided by the mean baseline power across all trials within a session. For the analysis of raw prestimulus power modulations no baseline correction was applied. We focused our analysis of EEG power modulations around target onsets on those electrodes that processed the visual stimulus. To this end, we averaged the power modulations or raw power across eleven occipito-parietal electrodes that showed stimulus-induced responses in the gamma-band range (59–100 Hz). See Kloosterman et al. [20] and Meindertsma et al. [21] for a similar procedure.

Condition-related EEG power modulation To test at which frequencies raw EEG power differed for the liberal and conservative conditions, we averaged power modulation from 0.8 s up to 0.2 s (i.e. up to half the window size used for spectral

analysis, to avoid contamination of post- with pre-stimulus activity [4]) from trial onset. Then, we expressed the power at each frequency in units of percent signal change with respect to the conservative condition and statistically tested whether this signal differed from zero (Figure 4D) (see Statistical comparisons).

Response gain model test To test the prediction of increased gain during liberal of the gain model, we first averaged activity in the 8–12 Hz range from 0.8 to 0.2 s before trial onset (staying half our window size from trial onset, to avoid mixing pre- and post-stimulus activity, also see Iemi et al. [4]) and took the log transform, yielding a single scalar value per trial expressing neural excitability. If this interval was not completely present in the trial due to preceding events (see Trial extraction), this period was shortened accordingly. Trials in which the scalar was > 3 standard deviations away from the participant's mean were excluded. We then sorted all single trials for each participant in ascending order of excitability and assigned them to ten equally-spaced bins ranging from the lowest to the highest excitability scalars present within that participant. Adjacent bin ranges overlapped for 50% to stabilize estimates (see Rajagovindan and Ding [8] for a similar procedure). Then we averaged the corresponding log-transformed gamma modulation of these trials (consisting of the average power within 59–100 Hz 0.2 to 0.6 s after trial onset) and normalized each participants response by subtracting the minimum gamma power during the conservative condition from all bins. Finally, we averaged across participants and plotted the excitability bin number against the normalized gamma power for each condition. To statistically test the gain prediction, we employed a three-way repeated measures ANOVA (see Statistical comparisons). For plotting purposes (Figure 4D), we computed within-subject error bars by removing within each participant the mean across conditions from the estimates.

Drift diffusion modeling We fitted the drift diffusion model to our behavioural data, for each subject individualy, and separately for the liberal and conservative conditions. We fitted the model using a *G* square method based on quantile RT's (RT cutoff, 200 ms, for details, see Ratcliff et al. [27]), using a tailored version of the HDDM 0.6.0 package [51] (code available at Github). The RT distributions for "yes" responses were represented by the 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles, and, along with the associated response proportions, contributed to *G* square. In addition, a single bin containing the number of "no" responses contributed to *G* square. Fitting the model to RT distributions for "yes" and "no" choices (termed 'stimulus coding' in Wiecki et al. [51]), as opposed to the more common fits of correct and incorrect choice RT's (termed "accuracy coding" in Wiecki et al. [51]), allowed us to estimate parameters that could have induced biases in subjects' behavior.

Parameter recovery simulations showed that letting both the the starting point of the accumulation process and drift criterion (an evidence-independent constant added to the drift toward one or the other bound) free to vary with experimental conditions is problematic for data with no overt "no" responses (data not shown). Thus, to test whether shifts in drift criterion or starting point underlied bias we fitted three separate models. In the first model ("basic model"), we allowed only the following parameters to vary between the liberal and conservative condition: (i) the mean drift rate across trials; (ii) the separation between both decision bounds (i.e., response caution); and (iii) the non-decision time (sum of the latencies for sensory encoding and motor execution of the choice). Additionaly, the bias parameters starting point and drift criterion were fixed with experimental condition. This model served as the baseline against which to compare the two models that could explain shifts in choice bias. The second model ("starting point model") was the same as the

basic model, except that we let the starting point of the accumulation process vary with experimental condition, whereas the drift criterion was kept fixed for both conditions. The third model ("drift criterion model") was the same as the basic model, except that we let the drift criterion vary with experimental condition, but the starting point was kept fixed for both conditions. We used Bayesian Information Criterion (BIC) to select the model which provided the best fit to the data [52]. The BIC compares models based on their maximized log-likelihood value, while penalizing for the number of parameters.

Statistical comparisons We used two-sided permutation tests (10,000 permutations) [53] to test the significance of behavioral effects and the model fits (Figures 1, 5). To quantify power modulations after (non-)target onset, we tested the overall power modulation for significant deviations from zero. For these tests, we used a cluster-based permutation procedure to correct for multiple comparisons [54]. For time-frequency representations of power modulation (Figure 2), this procedure was conducted across all time-frequency bins. For frequency spectra (Figure 3D), this procedure was performed across all frequency bins. To test whether there was any evidence for increased gain in the liberal compared to the conservative condition, we conducted a three-way repeated measures ANOVA (condition (conservative, liberal) x brain activity type (prestimulus alpha, poststimulus gamma power) x bin level (1–10)) using SPSS 23 (IBM, Inc.), inspecting linear and quadratic contrasts. As sphericity was violated in this model ($p < 0.0001$), we report both the uncorrected and Greenhouse-Geisser-corrected p-values.

Supplemental Information

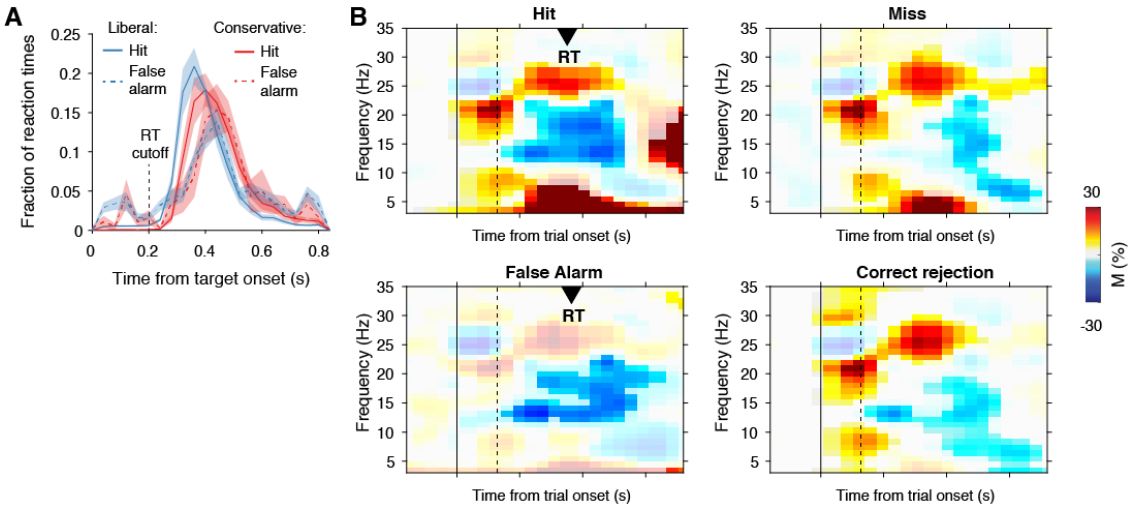
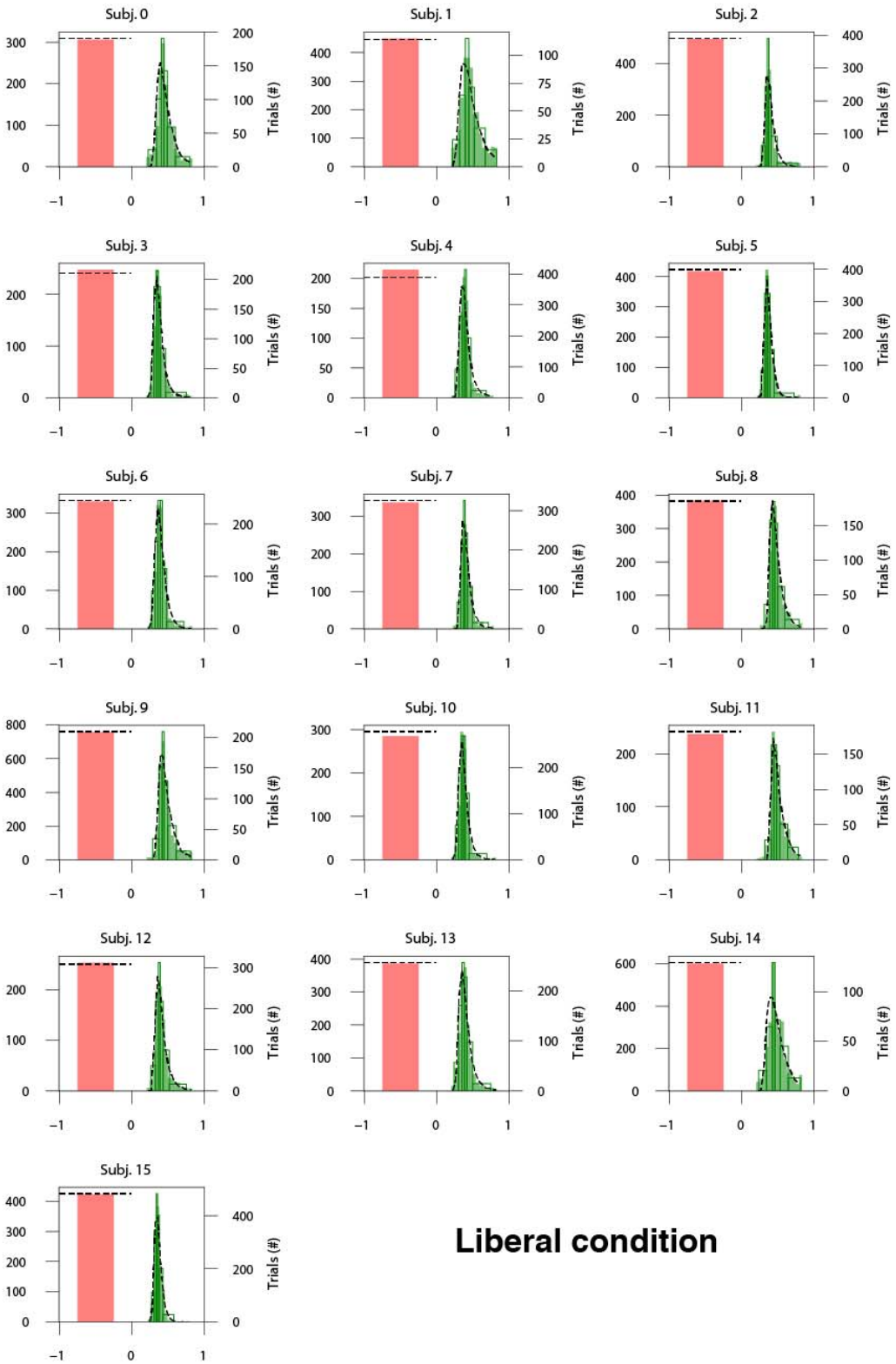
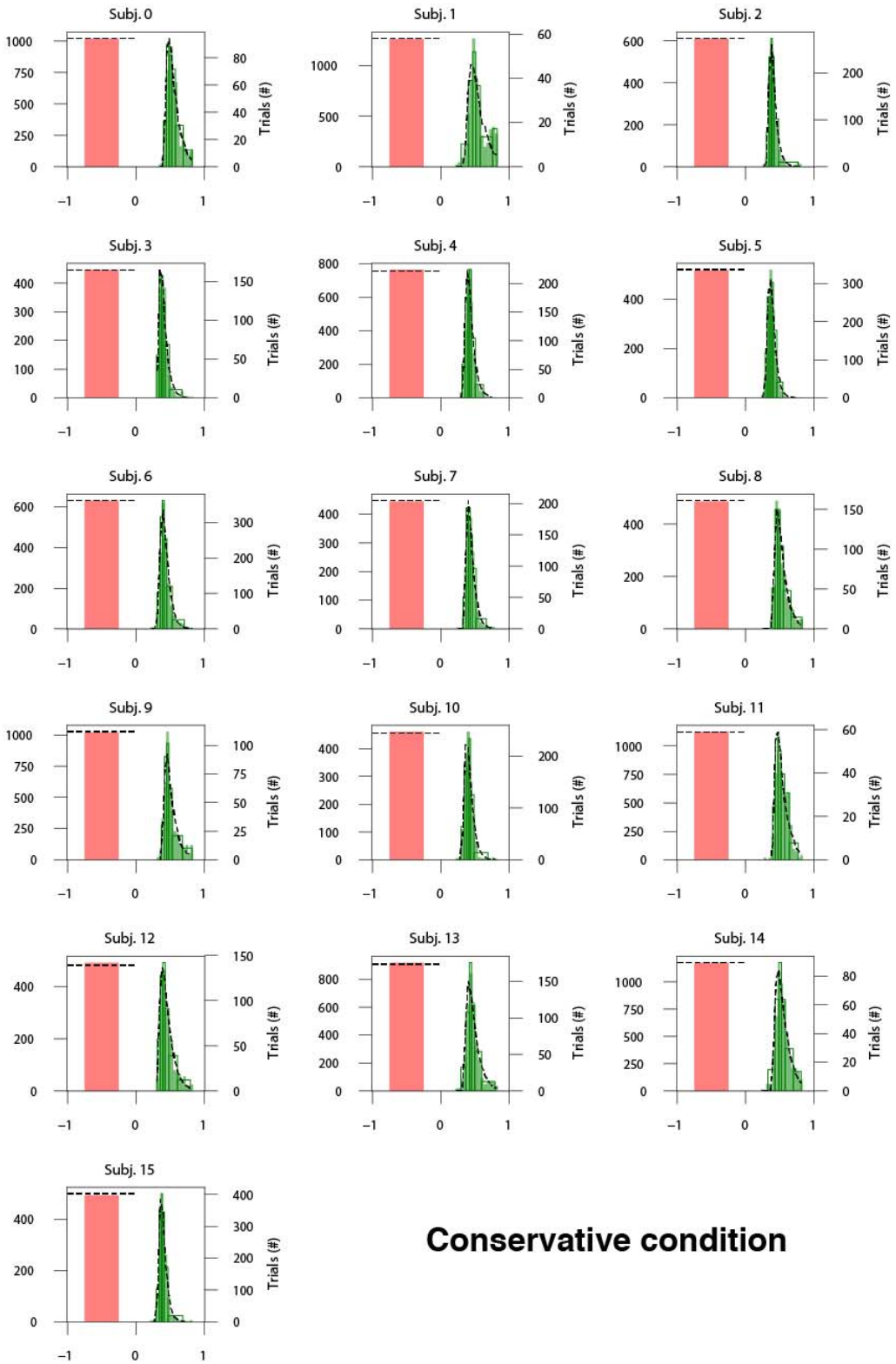


Figure S1 | Behavioral and neurophysiological evidence that participants were sensitive to the implicit task structure. Related to Figures 1 and 3. A. Subject-average RT distributions for hits and false alarms in both conditions, indicating that participants were sensitive to trial onset despite the fact that trial onsets were only implicitly signaled. **B.** Time-frequency representations of low-frequency EEG power modulations with respect to the prestimulus period (–0.4 – 0 s), pooled over the two conditions, indicating that participants detected the onset of a trial even when neither a target was presented nor a response was given (correct rejections). Saturated colors indicate clusters of significant modulation, cluster threshold $p < 0.05$, two-sided permutation test across participants, cluster-corrected; N = 15). Solid and dotted vertical lines respectively indicate the onset of the trial and the target stimulus. M, power modulation.



Liberal condition



Conservative condition

Figure S2 | Drift diffusion model fits for both conditions and for each participant. Related to Figure 5. Pink bars, number of “No” trials; Green bars, RT distribution for “Yes” trials; dotted lines, model fits for the drift criterion model.