Adaptive gain control in the human dorsal anterior cingulate cortex

Vickie Li[1*], Elizabeth Michael[2], Jan Balaguer[1], Santiago Herce Castañón[1] and Christopher Summerfield[1]

[1] Department of Experimental Psychology
University of Oxford
15 Parks Road
Oxford, UK

[2] Department of Psychology
University of Cambridge
Downing Street
Cambridge
CB2 3EB

*Corresponding author: chui.li@psy.ox.ac.uk

Abstract

When making decisions, humans are often distracted by irrelevant information. Distraction has different impact on perceptual, cognitive and value-guided choices, giving rise to well-described behavioural phenomena such as the tilt illusion, conflict adaptation, or economic decoy effects. However, a single, unified model that can account for all these phenomena has yet to emerge. Here, we offer one such account, based on adaptive gain control, and additionally show that it successfully predicts a range of counterintuitive new behavioural phenomena on variants of a classic cognitive paradigm, the Eriksen flanker task. We further report that BOLD signals in the dorsal anterior cingulate cortex (dACC), rather than reflecting the cost of resolving conflict among responses, index a gain-modulated decision variable predicted by the model. This work unifies the study of distraction across perceptual, cognitive and economic domains, and offers new insights into the puzzling pattern of neural responses typically observed in the dACC.

Keywords: cognitive control; fMRI; tilt illusion; conflict monitoring; gain control

Introduction

Decisions about sensory signals, cognitive propositions, or economic prospects are often made in the context of competing or distracting information. Consider the following everyday situations: you are judging whether a painting hangs straight on the wall, but the nearby pictures are hung askew; you are waiting at a red stop signal, but the car in front decides to jump the light; you are contemplating the purchase of a new watch, but it is displayed next to a range of more elegant but unaffordable models. In each of these situations, the best decisions will be made by ignoring the distracting sensory signals (the competing picture frames, vehicles, or watches) and focussing exclusively on the choice-relevant information. This normative contention can be formalised in a variety of ways, for example via the notion that rational choices should be independent of irrelevant alternatives (Luce, 1959; von Neumann and Morgenstern, 1944) or that sensory signals should be weighted lawfully by their reliability and relevance to the choice at hand (Kording, 2007; Ma et al., 2006; Savage, 1954; Wald and Wolfowitz, 1949).

Nevertheless, empirical observations suggest that humans decision-makers are unduly influenced by distracting information. Consider a generic problem in which a target stimulus X^i and distracters X^j occur at fixed locations *i* and *j*. In this general formulation, decision values *X* may be perceptual features (such as the tilt of a grating) or economic attributes (such as the quality of a consumer product) that are to be evaluated or categorised. Humans show systematic biases that reflect the influence of the distracters on decisions about the target. For example, vision scientists have long studied the "tilt illusion", in which the reported orientation of X^i (e.g. a central grating) is repulsed away from the mean tilt of X^j (surrounding gratings with similar but nonidentical tilt; **Fig. 1a**) (Blakemore et al., 1970). In cognitive psychology, the influence of distracter items is usually studied with a view to understanding the attentional or control mechanisms that allow information to be selected in the face of conflict. For example, in the classic Eriksen flanker task, observers classify a target stimulus (e.g. a central arrow) that is flanked by distracters (e.g. arrows pointing in compatible or incompatible directions) (Eriksen and Eriksen, 1974; Kopp et al., 1994). It is ubiquitously observed that incompatible flankers incur a cost, and compatible flankers a benefit, in response times (RTs) and accuracy relative to a neutral condition (**Fig. 1b**). Finally, behavioural and neural economists have charted the irrational influence that a decoy alternative of value X^j has on choices between two choice-relevant prospects X^{i+} and X^{i-}, where $X^{i+} > X^{i-}$ (Chau et al., 2014; Louie et al., 2013; Tsetsos et al., 2010; Tversky, 1972). A common finding is that rational choices (i.e. for $X^{i+} > X^{i-}$) initially decline as X^j increases in value, but then increase sharply as X^j comes to approximately match the other two items in value (**Fig. 1c**).

Figure 1

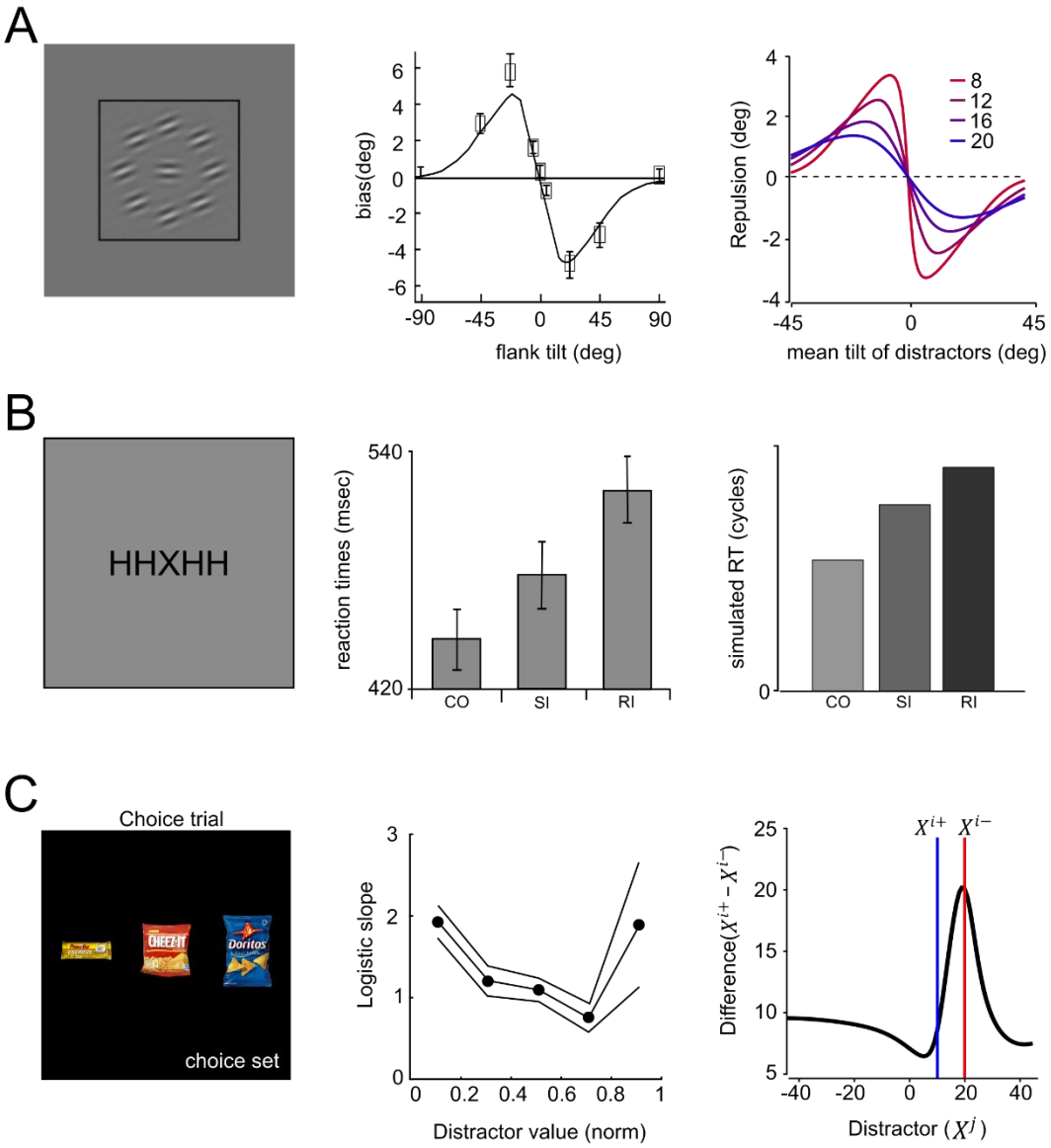

Figure 1. The effect distraction across perceptual, cognitive and economic domains. (A) Left panel: subjects were asked to discriminate the tilt (relative to horizontal) of a central Gabor surrounded by tilted distracters. Middle panel: participants are biased to report the target as more clockwise when the flankers were counterclockwise and vice versa (the "tilt illusion", middle panel; images adapted from Solomon & Morgan, 2006, permission pending). Right panel: simulation of the adaptive gain model recreates the tilt illusion and further predicts that the magnitude of the bias is modulated by flanker variance; coloured lines reflect flanker standard deviation from low (red) to high (blue). **(B)** Left panel: In the Eriksen flanker task, participants respond with a key press to a central letter while ignoring the flankers. Middle panel: response times are the fastest on CO ('congruent') trials, then the SI ('Stimulus Incongruent') trials, and slowest on RI ('Response incongruent') trials; Middle panel; adapted from van Veen & Carter, 2002, permission pending). See methods for details of how CO, RI and SI trials were defined. Right panel: the adaptive gain model predicts the same pattern of reaction time across the three conditions. **(C)** Left panel: participants chose the most preferred of three food items. Middle panel: increasing the value of the least-preferred item reduces the choice efficiency (i.e. probability of choosing the highest-valued target) as the normalised distractor value increases, shown by logistic slope from fitting logistic choice functions on humans choice. Using the adaptive gain model, we simulated the subjective difference between two tilts (blue and red line) as a function of a third distracting tilt (x-axis). The subjective difference is first reduced and the increased in a qualitatively similar fashion.

In the fields of psychology, economics, and neuroscience, diverse theoretical proposals have been offered to explain the cost that distracters incur during decision-making. These include models that describe how control systems detect and resolve conflict among inputs (Botvinick et al., 2001; Bundesen, 1998), accounts that emphasise inhibitory interactions among competing sensory neurons, or favour a normalisation of stimulus values by a local average or range (Heeger, 1992; Louie et al., 2013; Schwartz et al., 2009; Soltani et al., 2012), and Bayesian accounts that model spatial uncertainty among targets and distracters (Dayan and Solomon, 2010; Logan, 1996; Yu et al., 2009), or that assume strong priors on the compatibility of decision information (Yu et al., 2009). These accounts disagree about the computational mechanisms involved, the neural processing stages at which the cost of distraction arises, and about the brain structures that are recruited to protect decisions against irrelevant information. For example, divisive normalisation mechanisms may occur in sensory neurons in visual cortex (Heeger, 1992), or amongst value representations in the orbitofrontal cortex (Padoa-Schioppa, 2009), whereas the control systems that detect and resolve conflict have been attributed to medial and lateral prefrontal structures (Botvinick et al., 2001). As such, the field currently lacks a single, unified theory that can account for the effect of distraction on human decisions, or an integrated neural account of its implementation across perceptual, cognitive and economic domains.

The goal of the current paper is to offer such an account. We begin with a simple computational model that is motivated by past work showing that contextual signals determine the gain of processing of consistent (or “expected”) features during decision-making tasks (Cheadle et al., 2014). Our model, which is described here at the level of neural population codes, proposes that contextual signals sharpen the tuning curves of neurons with compatible preference for decision-relevant features, and is motivated by a large literature emphasising the need for adaptive gain control in the service of efficient coding (Rangel and Clithero, 2012; Simoncelli and Olshausen, 2001). Using computational simulations, we first show that the model can recreate qualitatively two classic phenomena in very different domains – perceptual choice (the tilt illusion) and economic choice (decoy effects). Next, we turn our attention to a task that has been a mainstay of cognitive studies of distraction – the Eriksen Flanker task. We built novel variants of the task in which the statistics of the flankers, and the difference between target and the decision bound can vary across conditions. Our simulations show that the model predicts a range of striking, counterintuitive behavioural findings, including “reverse” compatibility effects (where fully-visible, compatible flankers actually hinder, rather than help, behavioural performance). Over 4 behavioural experiments involving human participants, we validate these predictions, using visual stimuli defined by both tilt and colour. Finally, we use functional brain imaging to show that the modulatory influence on decision signals predicted by the model correlates with BOLD signals the dorsal anterior cingulate cortex (dACC) and interconnected structures, where neural signals have variously been implicated in the context-sensitive encoding of action values (Kolling et al., 2016b), and the expected value of cognitive control (Shenhav et al., 2016). We show how our framework, which is not wholly inconsistent with either account, can bring together diverse views concerning the function of this controversial brain region (Heilbronner and Hayden, 2016).

Results

Our adaptive gain model is based on a framework that was previously developed to understand how humans performing perceptual decision-tasks adapt to the context provided by information that is proximal in space and time (Cheadle et al., 2014; Michael et al., 2014). Inputs arrive at a population of n neurons each characterised by a Gaussian tuning curve centred on its preferred feature value θ_k. Each neuron k responds to the target stimulus X^i with rate $R_k = f\left(X^i|\theta_k, \sigma_k\right)$, where $f(X|\theta, \sigma)$ denotes the probability density function of the normal distribution with mean θ and variance σ^2.

The estimated output of the neural population is then decoded into a subjective percept of the target $\widehat{X^i}$ by weighting the population activity R by the corresponding feature values θ.

$$\widehat{X^i} = \sum_{k=1}^{n} R_k \cdot \theta_k$$

[eq.1]

When the gain is uniformly spread across the feature space (i.e. the tuning widths σ_k for all neurons are equal) this approach faithfully decodes each input to its original feature value. However, our model proposes that the context provided by the distracters modulates the sharpening of neuronal tuning, with a tuning width envelope that matches the inverse distribution of contextual features X^j with mean mX^j and standard deviation sX^j.

$$\sigma_k = \sigma^{max} - f\left(\theta_k|mX^j, sX^j + \varepsilon\right) \cdot n$$

[eq.2]

In other words, neurons with preferred orientation that matches mX^j have the sharpest tuning curves, and these tuning curves are even sharper if the flanker variance (sX^j) is low (see **Fig. 2a)**. In equation 2, σ^{max} denotes the maximum tuning width in the population, and ε is a noise parameter. In all model fitting procedures below, we allow ε to vary as a free parameter, recovering average values of ~5°, which concurs with previous accounts (Solomon and Morgan, 2006).

We first show how the model explains the both tilt illusion in perceptual choice tasks and decoy effects in economic choice tasks. In **Fig. 1a**, we plot the tilt bias over visual angle $\theta \in$ {-90,-89…90} predicted by the model as a difference of subjective and objective estimates of the target $\widehat{X^i} - X^i$ for different values of $mX^j \in$ {-45,-44…45} and $sX^j \in \{3, 7, 11, 15\}$. The model predicts that subjective estimates are repulsed away from the mean flanker value, as described in numerous previous studies (Blakemore et al., 1970; Gibson, 1937), and additionally that the strongest repulsion effect occurs when the flankers are consistent, i.e. are drawn from a distribution with low dispersion. This repulsion of the subjectively decoded values from their objective counterparts

occurs when the target feature close to, but is not identical to, the mean of the distracters (i.e. the location of sharpest tuning), because the variable tuning profile induces a skew in the population activity over features θ_k (see **Fig. S1** for a more detailed explanation). To model economic decoy effects, we envisage a choice between two prospects of value X^{i+} and X^{i-} (where $X^{i+} = 20$ and $X^{i-} = 10$ in arbitrary units) that is made in the context of distracters with average value X^j. We plot the difference in their corresponding subjective estimates $\hat{X}^{i+} - \hat{X}^{i-}$ as a function of mX^j, observing a pattern with a striking qualitative resemblance to that reported previously (Louie et al., 2013) (**Fig. 1b**). Again, the model's ability to predict this counterintuitive pattern comes from the repulsive effect induced by differential tuning across feature space. The model predicts that as the value of the decoy mX^j increases, repulsion is first strongest towards X^{i-} (leading to a reduced preference for the best option X^{i+}) but then, as the decoy approaches the two items in the choice set, repulsion is maximal for X^{i+}, reversing this effect. In further simulations, we systematically varied both the distance between X^{i+} and X^{i-}, and sX^j, we were able to capture the two-dimensional pattern of multi-alternative choice data described in a different study involving abstract shapes associated with different economic values (Chau et al., 2014) (**Fig. S2**).

Next, we used our model to simulate performance on a new variant of the flanker task that involves categorising a central grating X^i tilted at -45° from vertical, in the face of flanking gratings that are on average tilted in a compatible ($mX^j = -45°$) or incompatible ($mX^j = +45°$) fashion. **Fig. 2c** illustrates predictions from the adaptive gain model under two orthogonally varying factors: compatibility and flanker variability. Because flanker effects for fully-visible stimuli are strongest for response times, we plot the inverse model output $1/|\widehat{X^i}|$ as a proxy for RT (see eq.1). As can be seen, the model predicts a compatibility effect: faster RTs for trials where the target and distracters were of congruent sign. However, it also makes a new, testable prediction: that as sX^j (flanker variance) decreases, response times should be reduced on compatible trials but remain the same on incompatible trials (**Fig. 2c**). This occurs because on compatible trials (flankers at -45°), more gain is allocated to the target feature when the variance of the distribution of flanker orientations is lower. However, the gain allocated to incongruent targets (flankers at +45°) is negligibly different across different flanker variance levels since the neural gain they received are similar at the tail of the gain distribution, and so the model predicts that flanker variance should not affect performance on incongruent trials (**Fig. 2a**). By contrast, classic models propose that response conflict varies with the amount of crosstalk interference among responses (Botvinick et al., 2001). These models predict that heightened flanker variance should have equal impact on compatible and incompatible trials (see Methods for model details).

Figure 2

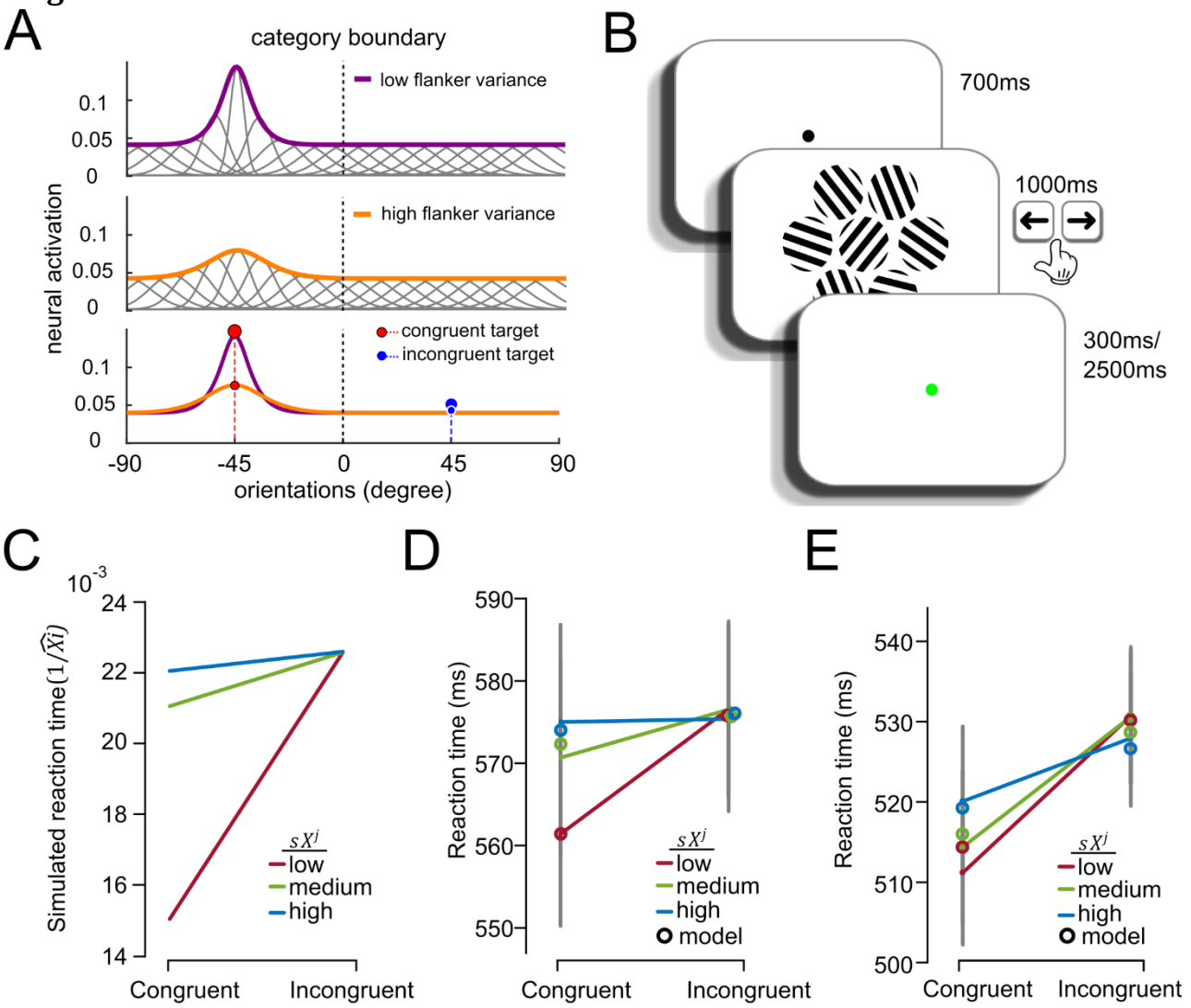

Figure 2. Adaptive gain model of the Eriksen flanker task. (A) Illustration of the adaptive gain model as applied to **Exp.1**. The model assumes that the population width tuning envelope is governed by the trial flanker mean (mX^j) and trial flanker variability (sX^j). Congruent targets (red circles) receive higher neural gain than incongruent targets (blue circles), leading to faster RTs. The model further predicts an interaction between congruency and flanker variability. Congruent targets receive high neural gain under low flanker variability trials (large circles) than high flanker variability trials (small circles) while incongruent targets under the two flanker variability trials received the same, low level of neural gain, meaning that flanker variability has no influence on RT for incongruent targets. **(B)** Experiment 1. Participant first saw a fixation dot, followed by an array that contained a central target and 6 surrounding flankers. They responded whether the central target was clockwise or counterclockwise to the vertical axis, receiving feedback after each response. **(C)** Simulation of RT ($1/|\widehat{X^i}|$) using the adaptive gain model for 3 levels of flanker variance (sX^j- in coloured lines) and congruency (x-axis). **(D - E)** Human reaction time pattern under three levels of flanker variance and congruency condition in **Exp. 1a** (Panel D) and **Exp.1b** (Panel E). Coloured circles represent the fitted mean reaction time from the adaptive gain model.

We tested this prediction in **Exp. 1** by asking healthy human participants to judge, relative to vertical, the tilt of a single target grating surrounded by 6 flanking distracter gratings (**Fig. 2b**). The orientation of the target grating X^i and the mean of the flankers mX_j were set to ±45° and the standard deviation of the flankers sX^j was varied at three levels: {0°, 15°, 30°} in **Exp. 1a** (n = 37) and {5°, 10°, 15°} in **Exp. 1b** (n = 36). In both experiments, we observed that flanker variance modulated RTs on compatible trials (**Exp. 1a**: $F_{1.91,68.6} = 9.041$, $p < 0.001$; **Exp. 1b**: $F_{1.91,68.86} = 20.16$, $p < 0.001$) but not on incompatible trials (both p-values ≥ 0.09). This finding was qualified by an interaction between compatibility and flanker variance (**Exp. 1a**: $F_{1.83,65.79} = 12.14$, $p < 0.001$; **Fig. 2d Exp. 1b**: $F_{1.95,70.06} = 20.32$, $p < 0.001$; **Fig. 2e**). Because previous work has shown that the ratio of compatible to incompatible flankers can modulate performance (Wu et al., 2016), we repeated this analysis limited to those trials where all flankers fell on the compatible/incompatible side of the boundary, finding a similar interaction for **Exp. 1a** ($F_{1.83,66.04} = 12.59$, $p < 0.001$) and **Exp. 1b** ($F_{1.65,57.66} = 15.77$, p < 0.001).

We fit our adaptive gain model to the data, and compared its predictions to those of a model proposing that RT depends on response conflict alone. The fits for the gain model (coloured circles) are shown superimposed upon the human data in **Fig. 2c and 2d**. We compared the models head-to-head by computing mean-squared error (MSE) in RT across conditions on half of the data (even trials), after estimating parameters from an independent dataset (odd trials). Bayesian model selection showed that the adaptive gain model fit the human data more closely than the conflict model, with exceedance probabilities for the adaptive gain model of 0.95 in **Exp. 1a** and 0.72 in **Exp. 1b**.

Next, we moved beyond the simple case in which X^i and mX^j fell equidistant the category boundary to a more complex design where they could vary independently around vertical at {±15°, ±30°, ±45°), and sX^j could once again vary at 3 levels {0°, 15°, 30°}. In this case, the model makes several predictions, some of them highly counterintuitive. Firstly, it predicts that there should be no main effect of congruence on RTs. In other words, the predicted inverse decision values $1/|\hat{X}_i|$ (or equivalently, unscaled RTs) are identical whether the target (X^i) and flanker mean mX^j are of the same and different sign. Secondly, the model predicts the existence of strong "reverse compatibility" effects under specific circumstances: there will be a disproportionate cost on congruent trials when the target X^i is closer to the category boundary than the mean of the flankers mX^j, e.g. when $X^i = 15$ and $mX^j = 30$ or $mX^j = 45$ (**Fig. 3a** upper left corner of each plot), and that this effect should diminish with increasing flanker variance (panels). Finally, although the model indicates that RTs will be dominated by the distance between X^i and the category boundary, it predicts that in the specific case where X^i and mX^j are both close to vertical, this cost will be strongly attenuated.

We tested these predictions using the flanker paradigm on two new cohorts of participants, one of which (**Exp. 2a,** n = 28) performed the tilt categorisation task described above, except with the full 6 (target mean) x 6 (flanker mean) x 3 (flanker variance) design. Another (**Exp. 2b**, n = 30) performed a task with the same design that involved judging the colour of a central circle (red vs. blue) surrounded by distracting flankers that varied continuously in colour from red to blue. Results from the two experiments were qualitatively very similar (see **Fig. S3, S4** for separate data) and so after normalising the feature values (tilt, colour) to an equivalent scale in the range [-1,1] we collapsed over them for display purposes. All three model predictions were strongly present in the human RT data (**Fig. 3a-c**). Firstly, splitting trials naively into compatible and incompatible, there was no overall significant difference in RT, either in the combined cohort (t_{57} = 1.14, p = 0.26) or separately for each experiment (**Exp. 2a**: p = 0.79; **Exp. 2b:** p = 0.17). Secondly, we see a strong cost on congruent trials when the flanker mean is further from the boundary than the target mean, as demonstrated by an $X^i \times mX^j$ interaction ($F_{3.72,208.23}$ = 17.48, p < 0.001) that was much weaker on incongruent trials ($F_{3.39,212.25}$ = 2.47, p = 0.049). This difference was qualified by a reliable three-way $|X^i| \times |mX^j| \times congruence$ interaction on human RTs ($F_{3.54,198.12}$ = 14.62, p < 0.001).

Figure 3

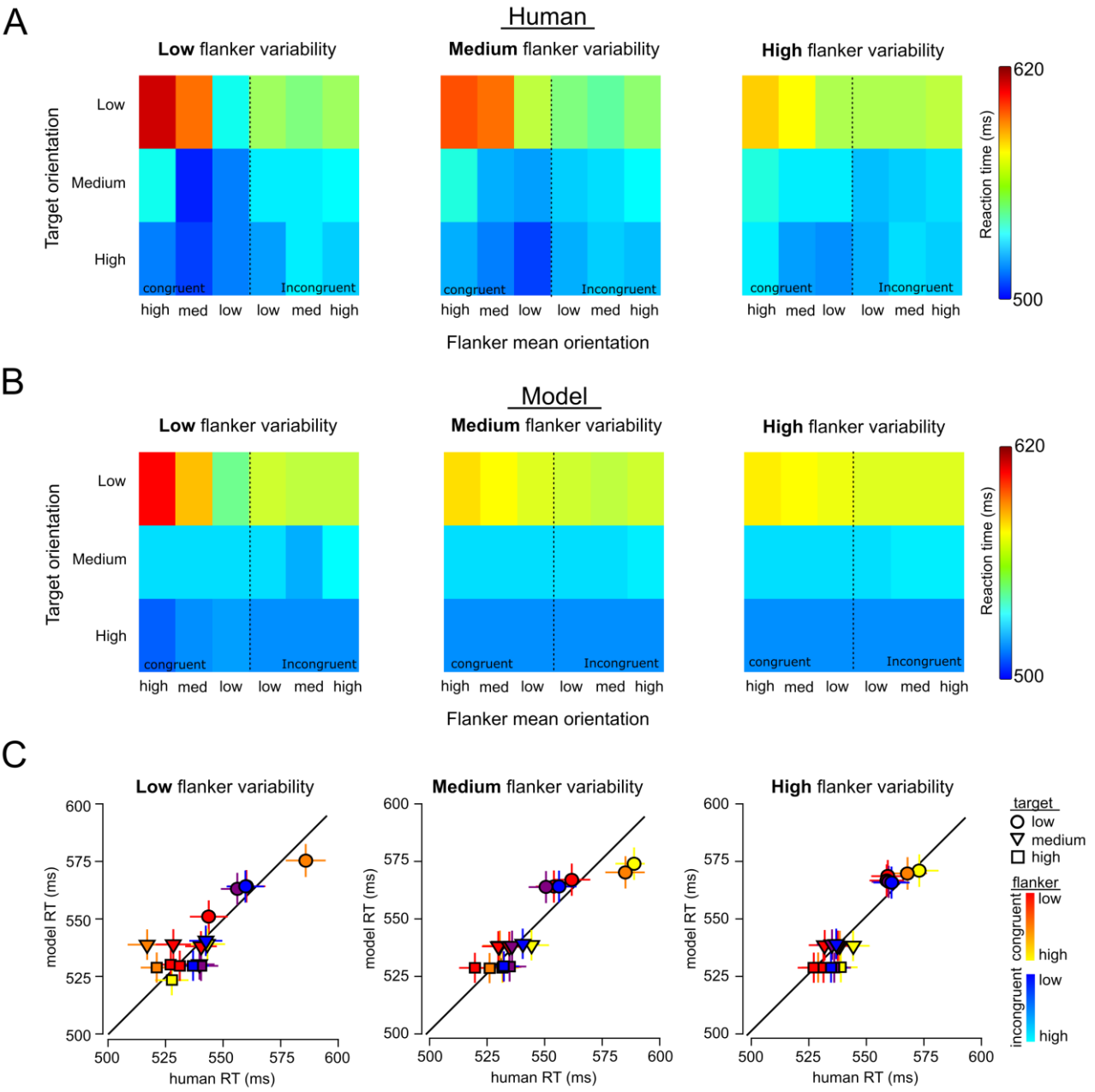

Figure 3. 4 way interactions (A) Surface plots showing the mean RT pattern in humans under different conditions (3 levels of target orientation x 3 levels of flanker mean orientation x congruency x 3 levels of flanker variability). Warmer colours correspond to longer reaction time. There is an overall cost when the target is close to the category boundary (top row of each surface plot). There is an additional cost when these targets were flanked by congruent flankers that are further from the boundary (top left corners of each subplot). This cost is radically reduced when the flankers are both congruent and close to the boundary (c.f. third most column, top row). The introduction of higher flanker variability reduces these additional costs in those conditions (overall faster RT across surface plots). **(B)** Fitted mean reaction time pattern from the adaptive gain model. **(C)** Mean (±SEM) RT in humans were cross-plotted against the fitted mean (±SEM) model RT for each condition. Warm colours (Red to yellow) correspond to levels of mean orientation from congruent flankers. Cold colours (Blue to cyan) correspond to levels mean orientation from incongruent flankers $|mX^j|$. Different shapes correspond to 3 levels of target decision variable $|X^i|$**.** Similar to upper plots, slower RT occurs specifically on trials where the target is close to the boundary and flankers are both congruent and have a mean orientation that is further from the category boundary.

The analyses described thus far pertain to RT data. In a final analysis of **Exp.2** we examined choices, using a previously-described approach based on probit regression to assess the weight (or influence) that distracters wielded over choices, as a function of whether their tilt was inlying or outlying with respect to category boundary (de Gardelle and Summerfield, 2011; Li et al., in revision). We attempted to predict choices on each trial as follows:

$$p(CW) = \Phi\left(\beta^0 + \beta^1 \cdot X^i + \beta_j^2 \cdot X^j\right)$$

[eq. 3]

Where X^j is tallied into 8 bins according to its signed feature value (e.g. from most counterclockwise to clockwise, or most red to blue). Plotting the coefficients β_j^2 associated with each bin j for model choices (obtained via the fitting procedure above), the model counterintuitively predicts that inlying flankers (those whose feature value falls close to the category boundary) will have a stronger impact on choices than outlying flankers. This occurs because the context-dependent gain field is, on average over all trials, Gaussian with a mean of approximately zero; in other words, the tuning of neurons near the category boundary is on average sharper. When we applied the same analysis to human data, we observed exactly the same pattern, with inlying flankers exerting a more distracting influence on choices (**Fig. 4a**). Applying the same regression model to predict RTs, the plots of β_j^2 for both humans and the model show the characteristic repulsion effect in choices, that is owing to the greater contextual modulation of neural tuning near the category boundary (**Fig. 4b**). Together, these analyses show that the human data resemble the counterintuitive model predictions over 4 different experimental datasets.

Figure 4

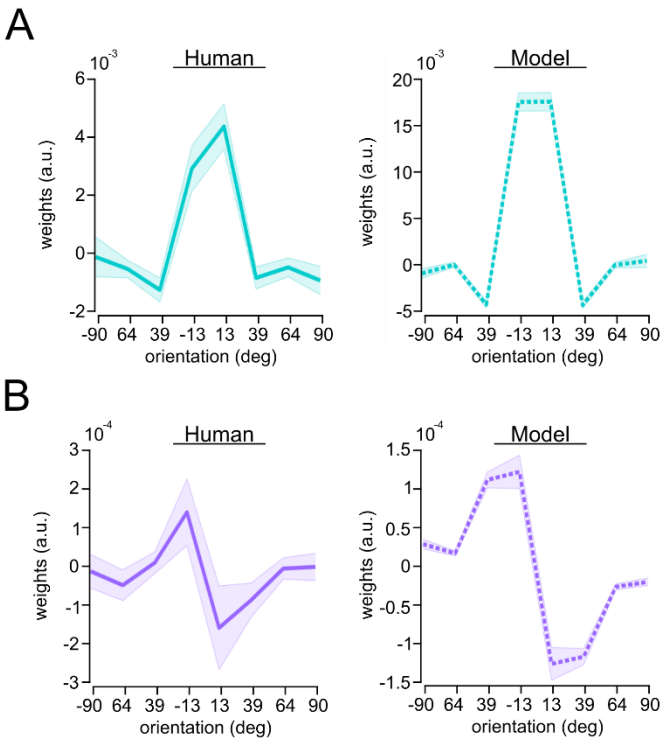

Figure 4. Choice bias driven by the flankers. Flanker tilts sorted into 8 bins from most CCW to most CW and used for predict choice (panel **A**; using probit regression) and response times (panel **B**) for both humans (left panels) and the best-fitting model (right panels). The x-axis shows the centre of each tilt bin, and the y-axis shows the resulting coefficients. The inverted-U-shape pattern in Panel A reveals that the inlying flankers have a stronger impact on choice than outliers for both humans and model. The response time data reveal the repulsive nature of this effect (Panel B solid purple line; note similarities with **Fig. 1a**). Higher absolute magnitude of beta weights imply that orientations that are close to the category boundary are more predictive of RT, similar to what was shown on choice in panel A. Shaded areas are the standard error of mean.

Functional brain imaging

Established theories propose that a brain network that prominently includes the dorsal anterior cingulate cortex (dACC) is involved in the recruitment of control processes that allow the brain to overcome distraction. Across a range of paradigms including the Eriksen Flanker task, the dACC responds with higher-amplitude BOLD signals on incompatible than compatible trials (Carter et al., 1998; Van Veen and Carter, 2002), and this effect is accentuated when the previous trial was compatible ('conflict adaptation'), as if the dACC is monitoring for conflict and signalling its onset (Kerns et al., 2004). However, the dACC is also implicated in decision processes more generally. For example, it signals the level of noise that corrupts an imperative stimulus during perceptual discrimination (Ho et al., 2009), and its proximity to a choice point or category boundary (Grinband et al., 2006); it responds to the relative economic value of an unchosen to a chosen option (Juechems et al., 2017; Tsetsos et al., 2014; Wunderlich et al., 2009), to the value of switching to a new task or context (Barber and Carter, 2004; Kolling et al., 2012), and to the update signals that occur as decision values change (Rushworth et al., 2011). The search for a unifying theory of the dACC has been one of the most challenging and controversial themes in cognitive neuroscience over recent years (Heilbronner and Hayden, 2016; Kolling et al., 2016a; Shahnazian and Holroyd, 2017; Shenhav et al., 2013).

To assess the role of the dACC and interconnected regions in adaptive gain control, we conducted a new experiment in which X^i, mX^j and sX^j varied parametrically from trial to trial, rather than in a conditionwise fashion. A new cohort of humans (**Exp.3**; n = 20) performed this task whilst we acquired BOLD signals from across the brain using functional magnetic resonance imaging (fMRI). Behavioural results of this experiment replicated those from **Exp.2** (**Fig. S5**), and so we focussed on neural analysis to test whether brain signals indexed decision information in a way that was predicted by the adaptive gain model. We began by confirming previous reports that the dACC responds more vigorously when a target feature lies closer to a category boundary, i.e. in our experiment, when the target orientation is closer to vertical (Grinband et al., 2006). We first regressed $|X^i|$ (i.e. proximity of the target to the category boundary) alone against BOLD signals occurring at the time of choice across the entire brain (GLM1). Consistent with previous observations, we observed a negative effect of $|X^i|$ in the dACC (peak: 2, 8, 54, t_{19} = 9.66, p_{fdr} < 0.001), as well as the anterior insula (AIC; peak: 34,34,2, t_{19} = 10.11, p_{fdr} < 0.001) and superior parietal lobe (SPL; peak: 18, -68, 58, t_{19} = 8.09, p_{fdr} < 0.001 see **Fig. 5a**). Extracting regions of

interest from these areas in a leave-one-out fashion across participants (see Methods), we then plotted how the BOLD signal varied in quartiles of both X^i and Xm^j (GLM2) and compared these signals to the predictions of (i) the adaptive gain model, (ii) an equivalent model with no adaptive gain, i.e. where all simulated cells had equivalent tuning width, and (iii) a model in which BOLD signals were driven by conflict alone (**Fig. 5b**). We found that the pattern of BOLD signals in all 3 regions closely resembled that predicted by the adaptive gain model, but not the other models (**Fig. 5c**). Specifically, although BOLD responses were elevated when the X^i was close to zero (dACC: $F_{1,19} = 52.37$, $p < 0.001$; AIC: $F_{1,19} = 53.4$, $p < 0.001$; SPL: $F_{1,19} = 48.94$, $p < 0.001$), this effect was exaggerated on those trials where mX^j was far from zero but of *compatible* sign (i.e. greater BOLD response in dACC, AIC and SPL on congruent relative to incongruent trials; dACC: $t_{19} = 3.03$, $p = 0.0069$; AIC: $t_{19} = 2.82$, $p = 0.011$; SPL: $t_{19} = 2.28$, $p = 0.034$). No such modulation was observed when X^i was far from zero, as predicted by the adaptive gain model.

Figure 5

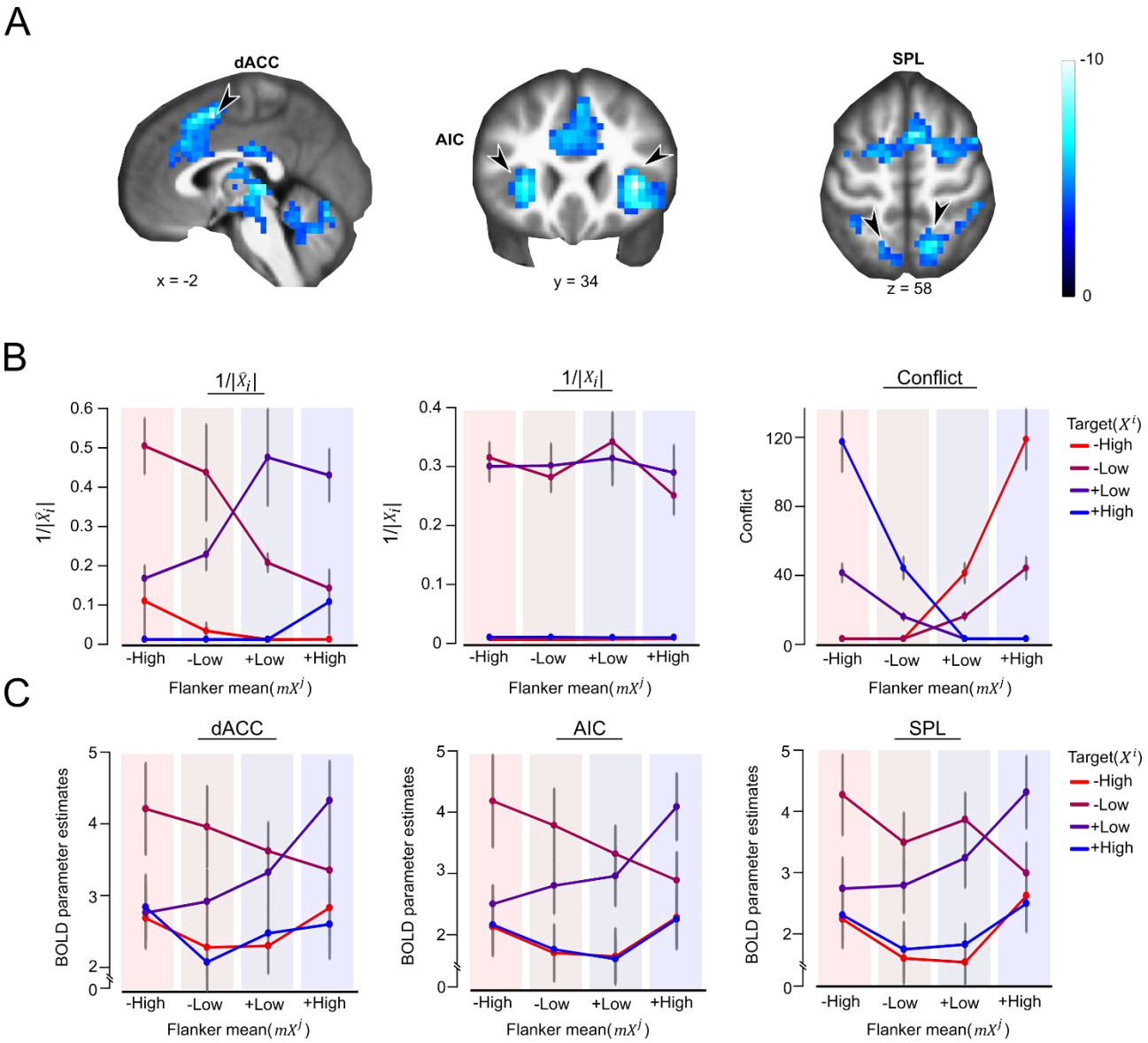

Figure 5. Effect of target orientation on BOLD signal (A) Brain areas correlating negatively with the absolute target decision variable, rendered onto a template brain in sagittal (left panel) and axial (middle and right) slices. Images were generated with an uncorrected threshold of p < 0.0001. **(B)** Parametric modulators X^i and mX^j were each being split into 4 quartiles (lines for X^i, x-axis or shaded area for mX^j). Mean predictions (±SEM) from different models: Reciprocal gain-modulated decision variable $1/|\hat{X}_i|$ (left panel). Reciprocal no-gain modulated target decision variable $1/|\hat{X}_i|$ (middle panel). Estimated conflict was plotted for from each quartile bins (right panel). Coloured lines correspond to 4 levels of $|X^i|$. Shaded coloured background (groupings on x-axis) corresponds to 4 levels of mX^j **(C)** Mean BOLD (±SEM) signal beta values for each level of $|X^i|$ and mX^j from the quartile bins. Left: dACC, Middle: AIC, and Right: SPL ROIs.

This suggests that a gain-modulated decision variable, rather than a conflict signal per se, is driving the dACC response. However, to quantify and compare the predictions of different models, we used Bayesian neural model comparison (Penny et al., 2006). We fit the adaptive gain model and the rival conflict models on the trial RTs. Model estimates $|X^i|$ and conflict computed in different ways (see Methods) from the best fitting parameters are then used to estimate BOLD signals. We computed, within the dACC, AIC and SPL, the posterior probability of the adaptive gain model conditioned on the BOLD signal, and compared the resulting estimates to those obtained for rival models. Exceedance probabilities strongly favoured the adaptive gain model over a decision model with no gain modulation, as well as over a family of conflict models (**Fig. 6a**). In other words, the dACC, along with AIC and SPL, code for a decision signal modulated in precisely the fashion predicted by the adaptive gain model.

Finally, we addressed a concern that dACC is simply exhibiting a BOLD signal that correlates with the response production time on each trial (Grinband et al., 2011). Disentangling these factors is challenging, because (as described above) the model does an excellent job of predicting RTs. Nevertheless, when we included both model output $1/|\hat{X}_i|$ and RTs as competitive predictors in the model (GLM4), we still recovered a significant activation in the dACC (peak: 2, 20, 50, t_{19} = 5.28,p_{fdr} = 0.023), and AIC (left peak: -30, 20, 6, t_{19} = 6.09 , p_{fdr} < 0.001; right peak: 34, 24, -6, t_{19} = 5.94, p_{fdr} < 0.001). In other words, the dACC BOLD signal correlates better with the demand predicted by the adaptive gain model than it does with time taken to produce a response on each trial.

How does our model explain previously-reported findings, such as the observation that the dACC responds to conflict (Botvinick et al., 1999), or to the relative value of an unchosen vs. a chosen option during economic choice (Boorman et al., 2013)? We have already shown that in the simple version of the Flanker task (c.f. **Exp.1a** where $sX^j = 0$), the model predicts a larger output $1/|\hat{X}_i|$ on incompatible relative to compatible trials. The model is thus in clear accord with a large literature indicating that dACC BOLD increases when target and distracters are incongruent in a simple version of the Flanker task (Hazeltine et al., 2000). We note that as described here, the adaptive gain model computes decision values independently on each successive trial, and thus in its current form would not predict conflict adaptation in the dACC. However, one could reasonably assume that adaptive effects may spill over from one trial to the next, i.e. that neural tuning width

will be partly modulated by the previous trial. Under this assumption, the adaptive gain model will successfully predict that responses should be faster on two successive incongruent or two successive congruent trials (Gratton et al., 1992), just as it successfully account for the observation that during categorisation of a multi-element array, response times are faster if the target array is preceded by a prime array with an equivalent level of feature variance (Michael et al., 2014).

However, we also note another facet of our results: that BOLD signals in the dACC, AIC and SPL ROIs correlate negatively with $|X^i|$ but positively with $|mX^j|$ (GLM5; **Fig. 6b**; dACC: t_{19} = 2.27, p = 0.035, AIC: t_{19} = 3.03, p = 0.003, SPL: t_{19} = 4.56, p = 0.007; this effect was also significant at the whole-brain level in voxels within the AIC and SPL, but not dACC). If we consider the target to be a "chosen" option and the flankers as a competing, "unchosen" option, the ensemble of findings reported here is reminiscent of the well-described finding by which dACC signals scale positively with the decision value associated with an unchosen option (i.e. the flankers) and negatively with the value of a chosen option (i.e. the target). Building on this intuition, we tested more directly the coding of model-predicted value of a chosen and unchosen option in a further simulation in which decision values for two stimuli were drawn randomly and independently from two distributions, and model output was converted to a choice via a softmax function (see Methods). This allowed us to correlate model output (i.e. predicted BOLD) with the value of the chosen and unchosen option, revealing a negative correlation with the former and a positive correlation with the latter, as previously reported (Boorman et al., 2009) (**Fig. 6c**). Our model thus unifies a number of disparate accounts that have emphasised a role for the dACC in tasks involving categorising perceptual stimuli, and choosing among economic prospects.

Figure 6

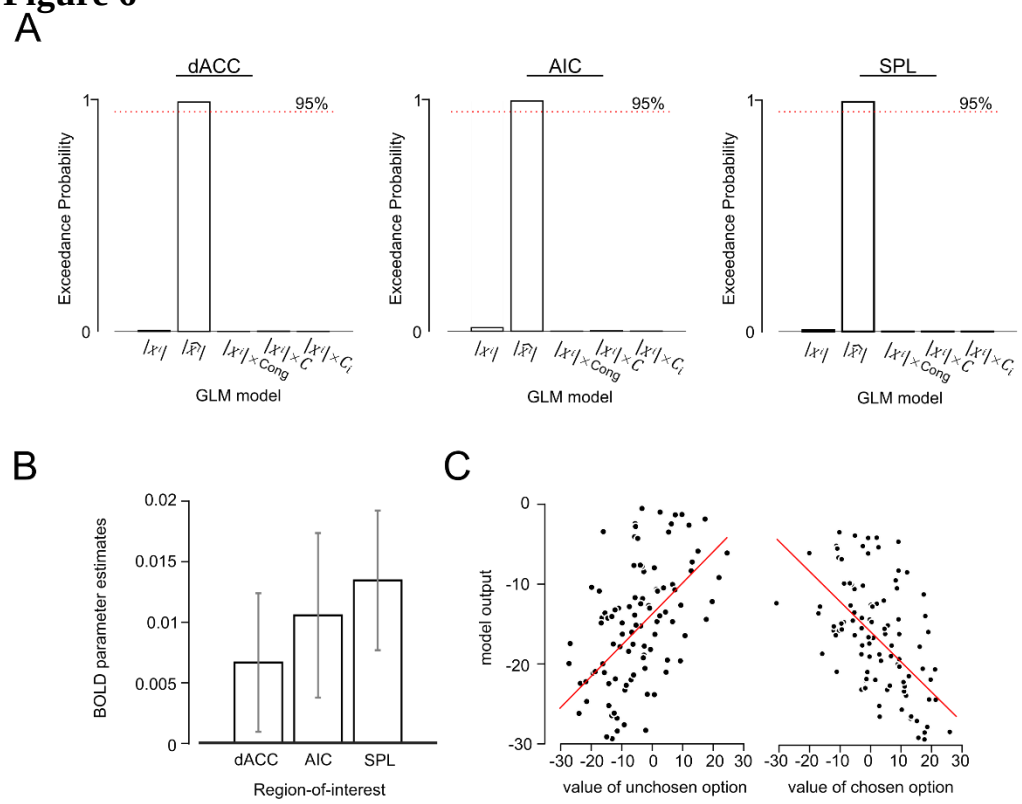

Figure 6. (A) Model comparisons across GLM models. Exceedance probability for each GLM and each ROI was computed using BMS. **(B)** ROI analysis on dACC, AIC, and SPL on the absolute flanker mean decision variable $|mX^j|$. Bar plots correspond to the mean (±SEM) beta estimates for this predictor from GLM5. **(C)** Simulated (negative) model output as a function of the value of chosen and unchosen option (see methods for details).

Discussion

Good choices are based solely on information that is relevant to the choice at hand, and rational agents will successfully ignore distracting signals when making decisions (Luce, 1959). However, across perceptual, cognitive and economic domains, human participants are observed to deviate from this rational principle. A range of different theories have been proposed to account for human susceptibility to distraction, but thus far, no single model has emerged that can account for phenomena as diverse as visual illusions, susceptibility to conflicting contextual features, or economic decoy effects. Here, we describe one such account. Our adaptive gain model has previously been shown to successfully account for diverse contextual influences in perceptual decision-making, including confirmatory biases in sequential sampling (Cheadle et al., 2014) and priming by second-order summary statistics in perceptual categorisation (Michael et al., 2014). Here, we show that not only can it account for classical effects of distraction across perceptual, cognitive and economic domains, but also that it successfully predicts a range of previously unreported, counterintuitive findings in a well-studied cognitive paradigm, the Eriksen flanker task (Eriksen and Eriksen, 1974).

The effect of distraction is most often modelled under the assumption that irrelevant features are imperfectly filtered during decision-making, driving residual activation that corrupts decisions. When target and distracters prompt conflicting sensorimotor responses, the resulting competition slows response times and increases error rates (Botvinick et al., 2001). A very successful neurocognitive theory proposes that dedicated processing systems have evolved in the primate medial prefrontal cortex that detect this conflict, and that are responsible for mobilising control mechanisms (associated with the lateral prefrontal cortex) to help mitigate the resulting costs (Egner and Hirsch, 2005). Here, we propose a differing view: one that emphasises the benefits of consistent context rather than the costs of inconsistent context. In the adaptive gain model, contextual features offer guidance as to where to best allocate gain across feature space, ensuring that neurons that code for the most prevalent (or “expected”) features have the sharpest tuning and thus provide the most sensitive outputs. The adaptive gain model is thus motivated by the more general view that the nervous system has evolved to code for efficiently for sensory inputs, reducing redundancy by dynamically adjusting the tuning properties of decision-relevant neurons to maximise sensitivity to expected features (Barlow, 1961; Simoncelli and Olshausen, 2001). In the flanker task, thus, the difference between compatible and incompatible trials arises at least in part because of a contextual facilitation mechanism at the decision level, akin to that described in sensory circuits (Stemmler et al., 1995), rather than because of an active cost of response competition. This idea is not without precedent in theories of control. In fact, the notion that a

flexible response to stimulus conflict is dependent on adaptive expectation mechanisms dates back to the original discovery of conflict adaptation by Gratton and colleagues more than 25 years ago (Gratton et al., 1992).

We take the opportunity to highlight two major features of our behavioural data that cannot be accounted for by standard accounts that emphasise the cost of conflict alone. Firstly, in **Exp.1a-b**, we found that a low-variance flanker array hastens response times on congruent trials, rather than prolonging response times on incongruent trials. This is consistent with an account that emphasises the benefit of consistent context rather than the cost of inconsistent context. Secondly, in **Exp.2a-b**, we observed that the longest response times were in fact observed on compatible trials, not incompatible trials. We replicated this finding across two different classes of visual feature: tilt and colour. According to our model, this cost occurred when the gain field dictated by the context repulsed the target subjectively closer to the category boundary, rendering choices more uncertain. Although such negative flanker effects have been reported with heavily masked stimuli, where they can be explained by differing timecourses of facilitatory versus inhibitory processes (Eimer, 1999), only rarely have such phenomena been reported for fully visible targets and distracters such as ours. Most interestingly, one such report occurred for a modified version of the flanker task where the targets were letters that were parametrically morphed between two possible identities, each corresponding to a possible flanker (Rouder and King, 2003). This report describes negative flanker effects when the target is most ambiguous, precisely parallelling our findings here for trials with small X^i and large mX^j, and a shift in the psychometric function that occurs with flanker identity in precisely the fashion predicted by our adaptive gain model (Cheadle et al., 2014).

Our behavioural findings were echoed in the neural data recorded from dACC, where BOLD signals were higher when targets fell closer to the category boundary, but these signals were positively modulated (yet higher) when the distracters mean was congruent but further from the boundary. Without further assumptions, a model based on conflict alone cannot account for these findings. We do not wish to argue that stimulus or response conflict do not ever incur an additional cost to accuracy and response times, or that such a cost is unable to drive the dACC. Nevertheless, in the current study, we found that such an account was not required to explain our data, and that a model embodying this assumption fit our data more poorly.

Our findings present a challenge to extant theories, but we acknowledge that our model is currently incomplete. A large literature implicates the dACC in the mechanisms by which we update the value of actions in dynamically changing environments (Behrens et al., 2007; Rushworth et al., 2011). Our experiments were conducted in stationary settings, and we do not doubt that these regions may play additional roles (potentially also related to gain control) when slower learning about a changing context is required. We also note an important shortcoming in our findings: we were unable to identify differing roles for the dACC, AIC and SPL, which seem to act as one in our study. We think it is likely that our BOLD data are simply indexing the output of a

decision process that involves modulation by distracting context, but are unable to make strong claims about the interim processes by which the computations proposed by the model occur. We suspect that exploring the role of adaptive gain control in dynamically-changing environments may shed more light on the differing contributions made by these regions, and we hope to pursue this question in future studies.

Author Contributions

V.L., E.M. and C.S. conceived the study and designed the experiments; V.L., J.B. and S.H.C. collected data., V.L. and C.S. analysed the data, J.B. contributed custom scripts for fMRI data analysis. V.L. and C.S. wrote the manuscript, with helpful input from the other authors.

Acknowledgements

We thank Tobias Egner for comments on the early version of the manuscript. This work was funded by European Research Council (ERC) Starter Grant (281628) to C.S., Croucher Foundation scholarship to V.L., Medical Research Council studentship to E.M. and the Wellcome Trust (0099741/Z/12/Z) to S.H.C.

References

Barber, A.D., and Carter, C.S. (2004). Cognitive control involved in overcoming prepotent response tendencies and switching between tasks. Cerebral Cortex *15*, 899-912.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In Sensory Communication (MIT Press).

Behrens, T.E., Woolrich, M.W., Walton, M.E., and Rushworth, M.F. (2007). Learning the value of information in an uncertain world. Nat Neurosci *10*, 1214-1221.

Blakemore, C., Carpenter, R.H., and Georgeson, M.A. (1970). Lateral inhibition between orientation detectors in the human visual system. Nature *228*, 37-39.

Boorman, E.D., Behrens, T.E., Woolrich, M.W., and Rushworth, M.F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron *62*, 733-743.

Boorman, E.D., Rushworth, M.F., and Behrens, T.E. (2013). Ventromedial Prefrontal and Anterior Cingulate Cortex Adopt Choice and Default Reference Frames during Sequential Multi-Alternative Choice. J Neurosci *33*, 2242-2253.

Botvinick, M., Nystrom, L.E., Fissell, K., Carter, C.S., and Cohen, J.D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature *402*, 179-181.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., and Cohen, J.D. (2001). Conflict monitoring and cognitive control. Psychological review *108*, 624-652.

Bundesen, C. (1998). A computational theory of visual attention. Philos Trans R Soc Lond B Biol Sci *353*, 1271-1281.

Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M., Noll, D., and Cohen, J.D. (1998). Science. Science *280*, 747-749.

Chau, B.K., Kolling, N., Hunt, L.T., Walton, M.E., and Rushworth, M.F. (2014). A neural mechanism underlying failure of optimal choice with multiple alternatives. Nat Neurosci *17*, 463-470.

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castanon, S., and Summerfield, C. (2014). Adaptive gain control during human perceptual choice. Neuron *81*, 1429-1441.

Dayan, P., and Solomon, J.A. (2010). Selective Bayes: attentional load and crowding. Vision Res *50*, 2248-2260.

de Gardelle, V., and Summerfield, C. (2011). Robust averaging during perceptual judgment. Proc Natl Acad Sci U S A *108*, 13341-13346.

Egner, T., and Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. Nat Neurosci *8*, 1784-1790.

Eimer, M. (1999). Facilitatory and inhibitory effects of masked prime stimuli on motor activation and behavioural performance. Acta Psychol (Amst) *101*, 293-313.

Eriksen, B.A., and Eriksen, C.W. (1974). Effects of noise letters upon identification of a target letter in a non- search task. Perception and Psychophysics *16*, 143-149.

Gibson, J.J. (1937). Adaptation with negative aftereffect. Psychological review *44*, 222-244.

Gratton, G., Coles, M.G., and Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. J Exp Psychol Gen *121*, 480-506.

Grinband, J., Hirsch, J., and Ferrera, V.P. (2006). A neural representation of categorization uncertainty in the human brain. Neuron *49*, 757-763.

Grinband, J., Savitskaya, J., Wager, T.D., Teichert, T., Ferrera, V.P., and Hirsch, J. (2011). The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. Neuroimage *57*, 303-311.

Hazeltine, E., Poldrack, R., and Gabrieli, J.D. (2000). Neural activation during response competition. J Cogn Neurosci *12 Suppl 2*, 118-129.
Heeger, D.J. (1992). Normalization of cell responses in cat striate cortex. Vis Neurosci *9*, 181-197.
Heilbronner, S.R., and Hayden, B.Y. (2016). Dorsal Anterior Cingulate Cortex: A Bottom-Up View. Annual review of neuroscience *39*, 149-170.
Ho, T.C., Brown, S., and Serences, J.T. (2009). Domain general mechanisms of perceptual decision making in human cortex. J Neurosci *29*, 8675-8687.
Juechems, K., Balaguer, J., Ruz, M., and Summerfield, C. (2017). Ventromedial Prefrontal Cortex Encodes a Latent Estimate of Cumulative Reward. Neuron *93*, 705-714.
Kerns, J.G., Cohen, J.D., MacDonald, A.W., Cho, R.Y., Stenger, V.A., and Carter, C.S. (2004). Anterior cingulate conflict monitoring and adjustments in control. Science *303*, 1023-1026.
Kolling, N., Behrens, T.E., Mars, R.B., and Rushworth, M.F. (2012). Neural mechanisms of foraging. Science *336*, 95-98.
Kolling, N., Behrens, T.E.J., Wittmann, M.K., and Rushworth, M.F.S. (2016a). Multiple signals in anterior cingulate cortex. Current opinion in neurobiology *37*, 36-43.
Kolling, N., Wittmann, M.K., Behrens, T.E., Boorman, E.D., Mars, R.B., and Rushworth, M.F. (2016b). Value, search, persistence and model updating in anterior cingulate cortex. Nat Neurosci *19*, 1280-1285.
Kopp, B., Mattler, U., and Rist, F. (1994). Selective attention and response competition in schizophrenic patients. Psychiatry Res *53*, 129-139.
Kording, K. (2007). Decision theory: what "should" the nervous system do? Science *318*, 606-610.
Li, V., Herce Castanon, S., Solomon, J.A., Vandormael, H., and Summerfield, C. (in revision). Robust averaging protects decisions from noise in neural computations.
Logan, G.D. (1996). The CODE theory of visual attention: an integration of space-based and object-based attention. Psychological review *103*, 603-649.
Louie, K., Khaw, M.W., and Glimcher, P.W. (2013). Normalization is a general neural mechanism for context-dependent decision making. Proc Natl Acad Sci U S A.
Luce, R.D. (1959). Individual Choice Behavior: A Theoretical Analysis (New York: Wiley).
Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat Neurosci *9*, 1432-1438.
Michael, E., de Gardelle, V., and Summerfield, C. (2014). Priming by the variability of visual information. Proc Natl Acad Sci U S A *111*, 7873-7878.
Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. J Neurosci *29*, 14004-14014.
Penny, W., Mattout, J., and Trujillo-Barreto, N. (2006). Bayesian model selection and averaging. In Statistical Parametric Mapping: The analysis of functional brain images, K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, eds. (London: Elsevier).
Rangel, A., and Clithero, J.A. (2012). Value normalization in decision making: theory and evidence. Current opinion in neurobiology *22*, 970-981.
Rouder, J.N., and King, J.W. (2003). Flanker and negative flanker effects in letter identification. Percept Psychophys *65*, 287-297.
Rushworth, M.F., Noonan, M.P., Boorman, E.D., Walton, M.E., and Behrens, T.E. (2011). Frontal cortex and reward-guided learning and decision-making. Neuron *70*, 1054-1069.
Savage, L.J. (1954). The Foundations of Statistics (New York: Wiley).
Schwartz, O., Sejnowski, T.J., and Dayan, P. (2009). Perceptual organization in the tilt illusion. J Vis *9*, 19 11-20.

Shahnazian, D., and Holroyd, C.B. (2017). Distributed representations of action sequences in anterior cingulate cortex: A recurrent neural network approach. Psychonomic Bulletin & Review, 1-20.

Shenhav, A., Botvinick, M.M., and Cohen, J.D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron *79*, 217-240.

Shenhav, A., Cohen, J.D., and Botvinick, M.M. (2016). Dorsal anterior cingulate cortex and the value of control. Nat Neurosci *19*, 1286-1291.

Simoncelli, E.P., and Olshausen, B.A. (2001). Natural image statistics and neural representation. Annual review of neuroscience *24*, 1193-1216.

Solomon, J.A., and Morgan, M.J. (2006). Stochastic re-calibration: contextual effects on perceived tilt. Proc Biol Sci *273*, 2681-2686.

Soltani, A., De Martino, B., and Camerer, C. (2012). A range-normalization model of context-dependent choice: a new model and evidence. PLoS Comput Biol *8*, e1002607.

Stemmler, M., Usher, M., and Niebur, E. (1995). Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics. Science *269*, 1877-1880.

Tsetsos, K., Usher, M., and Chater, N. (2010). Preference reversal in multiattribute choice. Psychological review *117*, 1275-1293.

Tsetsos, K., Wyart, V., Shorkey, S.P., and Summerfield, C. (2014). Neural mechanisms of economic commitment in the human medial prefrontal cortex. eLife *3*, e03701.

Tversky, A. (1972). Elimination by aspects: A theory of choice. Psychological review *79*, 281–299.

Van Veen, V., and Carter, C.S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. Journal of cognitive neuroscience *14*, 593-602.

von Neumann, J., and Morgenstern, O. (1944). Theory of Games and Economic Behavior (Princeton, NJ.: Princeton University Press).

Wald, A., and Wolfowitz, J. (1949). Bayes Solutions of Sequential Decision Problems. Proc Natl Acad Sci U S A *35*, 99-102.

Wu, T., Dufford, A.J., Mackie, M.A., Egan, L.J., and Fan, J. (2016). The Capacity of Cognitive Control Estimated from a Perceptual Decision Making Task. Scientific Reports *6*.

Wunderlich, K., Rangel, A., and O'Doherty, J.P. (2009). Neural computations underlying action-based decision making in the human brain. Proceedings of the National Academy of Sciences *106*, 17199-17204.

Yu, A.J., Dayan, P., and Cohen, J.D. (2009). Dynamics of attentional selection under conflict: toward a rational Bayesian account. J Exp Psychol Hum Percept Perform *35*, 700-717.

Methods

Subject details

For behavioural studies **Exp.1** and **2**, human participants were recruited via the online testing platform provided by Amazon Mechanical Turk (**Exp. 1a**: n = 37, 19 men, mean age: 31.8 years, SD = 8.09; **Exp.1b**: n = 36, 22 men, mean age: 31.12 years, SD = 9.57; **Exp. 2a**: n = 28, 14 men, mean age: 33.5 years, SD = 13.09; **Exp.2b**: n = 30, 16 men, mean age: 33.3 years, SD = 9.5). Mind here that the age information was collected by sets of age (e.g. 18-20, 21-30, 31-40 etc.) due to online testing, thus the reported age from Exp. 1a,1b,2a,2b was computed by taking the bin center of each age bin as the approximation of participants' true age. For **Exp.3**, 20 healthy volunteers (8 men, mean age: 24.25 years, SD = 4.61) with normal or corrected-to-normal vision and no history of neurological disorders were recruited to participate. All participants gave informed consent to participate in the study and were compensated at a rate of $6 per hour for **Exp. 1** and **2** and €10 per hour for the fMRI scanning session. All experiments were conducted in accordance with local ethical guidelines.

Task details

Experiment 1a, 1b and 2a. Each trial began with a fixation dot 700ms, followed by the simultaneous presentation of target and flankers for 1000 ms. Target and flankers were full-contrast square-waved gratings with a 180° phase (5 cycles, and 75 pixels in diameter). The flankers are presented within a circular aperture of the target grating with 1.5 pixels separation with adjacent flankers**.** The orientation of each individual grating in the flanker surround was chosen form a Gaussian distribution with a particular mean (mX^j) and standard deviation (sX^j; see main text for precise values). To ensure that the sampled orientations matched the expected distribution with the given mX^j and sX^j, resampling of orientation values occurred until the mean and standard deviation of orientations fell within 1° tolerance of the desired mX^j and sX^j. Within the stimulus presentation duration, participants had to indicate whether the central grating was clockwise or counterclockwise to the vertical axis by pressing the right and left keyboard arrows respectively. They were instructed to ignore the flanking gratings. Visual feedback was given immediately after their response. Participants completed 6 blocks of 96 trials for **Exp. 1a**, 10 blocks of 96 trials for **Exp. 1b**, and 10 blocks of 108 trials for **Exp. 2a**).

Experiment 2b. While keeping all other aspects of the experiment (e.g. timing) the same, in **Exp.2b** we substituted the grating stimuli with coloured stimuli from a continuous colour space. Target and flankers were circles (100 pixels diameter per circle; 19 pixels separation with adjacent flanking circles) with hue ranging from blue to red. Participants were asked to indicate whether the central colour circle was more red (right arrow) or blue (Left arrow) while ignoring the surrounding coloured circles. Individual colour values (*c*) were defined by RGB values using expression: RGB = [*c*, 0, 1-*c*] with the category boundary being defined as *c* = 0.5 or RGB = [0.5, 0, 0.5]. The colour value for the central target stimulus and the flanker mean colour value were drawn from the set {0.35, 0.4, 0.45, 0.55, 0.6, 0.65}. To facilitate comparison between **Exp.2a** and **Exp.2b**, we recentered the category boundary value to 0, which means the central target colour values (X^i) can be redescribed as ±0.05, 0.1, 0.15 with respect to a category boundary of zero. In this space individual flanker values were drawn from Gaussian distribution with means (mX^j) of ±0.05, 0.1, 0.15 and standard deviations (sX^j) of 0, 0.1, 0.2. Resampling of flanker feature values occurred until mX^j and sX^j fell within 0.01 from the desired values. Participants completed 10 blocks of 96 trials for this experiment.

Experiment 3 (fMRI study). All stimuli were presented on a grey background. Participants first saw a fixation dot for 500ms, followed by the presentation of stimuli for 500ms, and were asked to respond within 1500 ms of stimulus onset. Stimuli were tilted gratings as for **Exp.1** and **Exp.2a**. At the end of response window, visual feedback in the form of a green or red dot indicated whether the response was correct or not. There was a jittered interval of between 2-6 s (on average 4s) interposed between trials. Target tilt and flanker tilt statistics were varied independently on a trial-by-trial basis. On each trial, tilts of the target grating and flanker mean orientation were drawn independently from uniform distributions with a limit of -45° and +45°. Similarly, flanker standard deviation could be values between 0 to 30° from a uniform distribution. Resampling was used to ensure that the number of congruent and incongruent trials was matched across the entire experiment. Participants completed 4 (n = 10) or 5 (n = 10) blocks of the task in the scanner, with 130 trials on each block.

fMRI acquisition and preprocessing. Images were acquired with 3-Tesla Siemens TRIO with a 32-channel head coil using a standard echo-planar imagining sequence. Images were 64 × 64 ×32 volumes with voxel size 3.5×3.5×3.5 mm; acquired with a 2s repetition time and echo time of 30 ms. We acquired fMRI data in 4 or 5 runs with 425 volumes per run. All preprocessing and fMRI analyses were carried out using SPM12. Preprocessing of the imaging data includes realignment of function images, co-registration of anatomical scan to the mean functional image, followed by segmentation and spatial normalisation to the standard template brain of the Montreal Neurological Institute (MNI brain). Lastly function images were spatially smooth with a 6-mm full width half maximum Gaussian kernel. A 128-s temporal high-pass filter was applied to exclude low-frequency artefacts.

Design and behavioural analysis

For experiments 1 and 2, the design orthogonalised the manipulation of target feature value (X^i), mean of flankers (mX^j) and variability of flankers (sX^j). We can further designate trials as "congruent" when X^i has the same sign as mX^j or "incongruent" when X^i has the opposite sign as mX^j. In Experiment 1, we thus have a 3 × 2 (flanker variability × congruency) within-participant factorial design. In experiment 2, we introduced three levels of $|mX^j|$ and three levels of $|sX^j|$, resulting in 3 × 3 × 3 × 2 ($|X^i|$× $|mX^j|$ × sX^j × congruency) within-participant factorial design with 54 conditions. Full list of X^i, $|mX^j|$, and sX^j levels were displayed in **Table S1**. For **Exp. 1** and **2**, ANOVAS with Greenhouse-Geisser correction for sphericity were carried out at group-level analyses. A threshold of $p < 0.05$ was adopted for all behavioural analyses. We only analysed RT from correct trials, and additionally excluded trials where RT was faster than the 1% percentile or slower than the 99% percentile of the RT distribution. We used the same exclusion criteria across experiments.

Computational modelling

Adaptive gain model. The computations that describe the population coding version of the adaptive gain model are described in the main text. To fit model outputs to human response time data (i.e. on a common scale in ms), for each parameterisation we regressed inverse decision values against each individual participants' response times:

$$RT_{gain} = \beta^0 + \beta^1 \cdot 1/\left|\hat{X}_i\right|$$

[eq. 4]

This calculation of RT is equivalent to modelling the data with ballistic (noiseless) diffusion process, with the two additional parameters β^0 and β^1 encoding the drift rate and non-decision time respectively (fixed across conditions). Searching , exhaustively across values of σ^{max} and ξ , we identified the parameters that minimised mean squared error (MSE) between the human and model-predicted average RTs for each condition.

Conflict models. For the conflict model, we use a formulation described previously, whereby conflict C depends on the weighted product of competing inputs for the two actions A^{CW} and A^{CCW}

$$C = [A^{CW} \cdot A^{CCW}]$$

$$A^{CW} = g(mX^j) \cdot (1 - w) + g(X^i) \cdot w$$
$$A^{CCW} = h(mX^j) \cdot (1 - w) + h(X^i) \cdot w$$

[eq. 5]

where $g(X)$ and $h(X)$ respectively denote positive and negative linear rectification functions. In conflict model 1, activation for the two actions A^{CW} and A^{CCW} is proportional to the tilt of the target and flanker mean. Alternatively, it can also be defined according to the tile of the target and each individual flanker (conflict model 2).

$$C_i = [A^{CW}{}_i \cdot A^{CCW}{}_i]$$

$$A^{CW} = \sum_{1}^{j} g(X^j) \cdot \frac{(1 - w)}{6} + g(X^i) \cdot w$$

$$A^{CCW} = \sum_{1}^{j} h(X^j) \cdot \frac{(1 - w)}{6} + h(X^i) \cdot w$$

[eq. 6]

Where j is the number of flankers that are either congruent or incongruent. In other words, these models made different assumptions about conflict: that it was computed at the level of individual flankers (conflict model 2) or at the level of summary statistics (conflict model 1).

Finally, we compute RTs in a similar fashion as for the gain model:

$$RT_{conflict} = \beta^0 + \beta^1 \cdot C$$

[eq. 7]

To reduce the risk of overfitting we used cross-fitting, estimating model parameters on half the trials and computing mean squared error (MSE) for the other half. These MSE were then fed into Bayesian Model Selection (BMS) to compute exceedance probabilities.

Simulations of qualitative effects

When fit to human data, the adaptive gain model contained 2 free parameters: maximum tuning width (σ^{max}) and sensory noise (ε). Simulations of the model aimed at qualitatively recreating effects from the past literature (e.g. for **Fig.1**, **Fig. 6**) assumed a fixed σ^{max} ($\sigma^{max} = 10$) and a fixed ε (ε = 5) unless noted otherwise.

Tilt illusion. In this simulation, we plot the difference between the true target angle (here, zero) and the gain modulated decision value ($\widehat{X^i}$) as a function of flanker mean decision value ($mX^j \in \{-45, -44, \dots, +45°\}$) and flanker standard deviation ($sX^j \in \{3,7,11,15\}$). For each variant of flanker mean decision values and flanker standard deviation, $\widehat{X^i}$ is computed using equation 1 and 2. We then plot $\widehat{X^i}$ against levels of flanker mean decision value and flanker standard deviation in **Fig. 1a**.

Conflict effects. We computed a proxy of RT ($1/\widehat{X^i}$) for the 3 conditions: 'CO': Congruent, where the target shares the same response association with the flankers; 'SI': Stimulus Incongruent, where the target is perceptually different to the flankers but the response associations of the two are still the same; 'RI': Response Incongruent, where the target has a different response association as the flankers (**Fig. 1b**). In the simulation, flanker standard deviation sX^j is set to be 0 in both 'CO' and 'RI' condition (i.e. we assumed flanker standard deviation was ε, or 5°). In 'CO', X^i is equal to mX^j (both are +45°). In 'RI', mX^j has the opposite sign to the X^i. Lastly, we simulated 'SI' condition by assuming sX^j is higher than 0 i.e. $sX^j = 5$; individual flankers are variable but mX^j remained the same as X^i. We assumed a higher maximum tuning width ($\sigma^{max} = 15$) in this simulation.

Multi-alternative valued-based decision making task. In **Fig. 1c**, we simulated the difference between the model estimated decision values from the two targets ($X^{i+} = 20$ & $X^{i-} = 10$) as a function of a third distractors' orientation ($X^j \in \{-45, -44, \dots, +45°\}$). sX^j was assumed to be 5 in this simulation. In **Fig. S1**, we computed the model output associated with the highest-valued choice-relevant alternative ($\widehat{X^{i+}}$) and the next best alternative ($\widehat{X^{i-}}$), assuming that the mode of the gain field determined by the statistics lowest-valued (i.e. irrelevant) alternative X^j. We then plotted a quantity proportional to choice probability, $\widehat{X^{i+}} - \widehat{X^{i-}}$, i.e. the relative difference in model output for the best and next best options.

Value of chosen vs. unchosen option. We simulated the model output as a function of the value of a theoretical chosen and unchosen option. On each trial, decision values for two stimuli (X^1 and X^2) were drawn independently from two zero-mean Gaussian distribution with a standard deviation of 10. On every trial, we allowed simultaneous evaluation of each stimulus in the context of the other, i.e. we passed each stimulus through the model as target with the alternative as distracter. We then assumed that participants chose according to the relative subjective (i.e. model output) value of the maximum and minimum resulting values, using a value of 5 for the slope of the choice function:

$$D = \hat{X}^{max} - \hat{X}^{min}$$

$$CP = \frac{1}{1 + e^{\frac{-D}{5}}}$$

[Eq. 8]

This allowed us to plot the relationship between X and model output D separately for the chosen and unchosen options.

Behavioural model comparison. We implemented Bayesian model selection at the group level to obtain posterior likelihoods and exceedance probabilities as described by Stephan and colleagues(Stephan et al., 2009).

fMRI data analyses

Univariate Analyses. We analysed our data using Statistic Parametric Mapping (SPM12) with the general linear model (GLM) framework and in-house scripts running in Matlab. For all analyses, we ensured that sequential orthogonalisation of predictors in SPM was disabled. All GLMs also included regressors encoding the estimated movement parameters from preprocessing as a nuisance covariate. We modelled correct trials and trials that RT falls within 1 and 99 percentile distribution. All other trials were modelled separately as a nuisance regressor in all GLMs. We first constructed GLM1 with a single predictor encoding the parametric target decision values $|X^i|$of the stimulus, time-locked to the onset of the stimulus. We identified voxels activated negatively by this regressor to define regions of interest in the dACC, AIC and SPL. Activations in these regions survived FDR correction for multiple comparisons at $p < 0.05$. To avoid double-dipping for later analyses, each region was identified in a leave-one-out fashion, with each participant in turn being omitted from a group-level analysis, which was used to define a Region-of-Interest (ROI) with a threshold of uncorrected $p < 0.0001$, from which beta values were extracted from the left out participant. For GLM2, we discretise each parametric modulator (i) target decision values $|X^i|$and (ii) flanker mean feature value $|mX^j|$into 4 quartiles bins and included a total of 4 x 4 = 16 regressors (corresponding to each quartile bins of feature values) in the GLM. BOLD betas from each subject are extracted using each ROI masks defined by the leave-one-out analysis. In GLM3, we entered the adaptive gain model estimates $1/|\hat{X}_i|$ and binary regressor (congruency: +1 for congruent trials, -1 for incongruent trials) as competitive regressors. In GLM4, we replaced congruency with response time as a competing regressor to $1/|\hat{X}_i|$. In GLM5, we included the following predictors as parametric modulators of the stimulus at the time of the decision (i) target decision values $|X^i|$, (ii) flanker mean feature value $|mX^j|$, (iii) flanker precision, i.e. inverse flanker variability $1/sX^j$, (iv) absolute distance between target and flanker mean orientations $|X^i - mX^j|$, (v) the interaction of the distance between target and flanker mean with target decision values $|X^i - mX^j| \times |X^i|$, and lastly (vi) the interaction of the distance between target and flanker mean, target decision values and flanker variability ($|X^i - mX^j| \times |mX^j| \times 1/sX^j$). Full details of active voxels associated with these regressor were displayed in **Table S2-4**. Subsequent statistical tests on the betas extracted from each ROI were conducted using two-tailed one-sample *t*-test at a significance threshold of $p < 0.05$.

Neural Bayesian model comparison. We estimated the following GLMs using Bayesian statistics and conducted model comparisons to quantify and compare the predictions of different GLMs: (i) no-gain modulated decision value $|X^i|$ i.e. GLM1 described above, (ii) gain modulated decision value ($|\hat{X}_i|$; GLM6) , 3 GLMs that contained the no-gain modulated decision value and conflict defined in three ways (iii) GLM7: a binary congruency regressor (iv) GLM8: conflict computed using

equation 4 (v) GLM9: conflict computed using equation 5. Each GLM yielded log evidence, which then was then used to compute exceedance probabilities for each GLM within each masks from the leave-one-out analysis in GLM1.

Supplementary Information

Figure S1

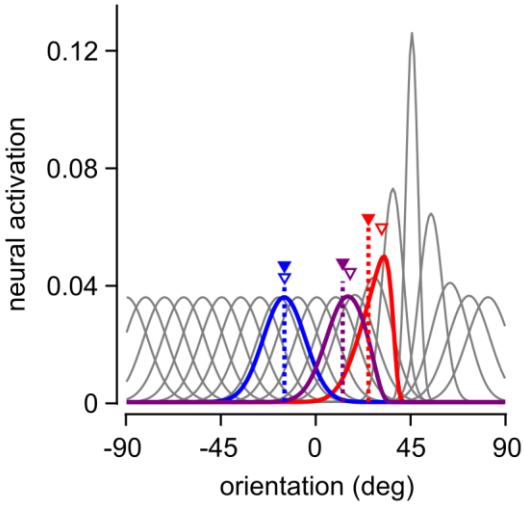

Figure S1. Tilt illusion prediction from the adaptive gain model. The illustration displayed the population activity vector θ_k for 3 hypothetical targets ($X^1 = -15°$, blue solid line; $X^2 = 15°$, purple solid line; $X^3 = 30°$, red solid line) from a population neurons that their tuning width envelope is based on a mX^j = 45° and sX^j = 5. The population activity for X^3 is more skewed than $X^1$and X^2, the hypothetical target which has the greatest distance to mX^j. This target is now perceived as closer to the category boundary at 0°. The strength of repulsion away from $mX^j$is also the greatest with this target. This can be seen clearly by comparing filled coloured triangles and dotted line (decoded orientation from the model) against unfilled coloured triangles (objective orientation). The difference between decoded orientation and objective orientation is the greatest for X^3 (red triangles) and smallest for X^1(blue triangles).

Figure S2

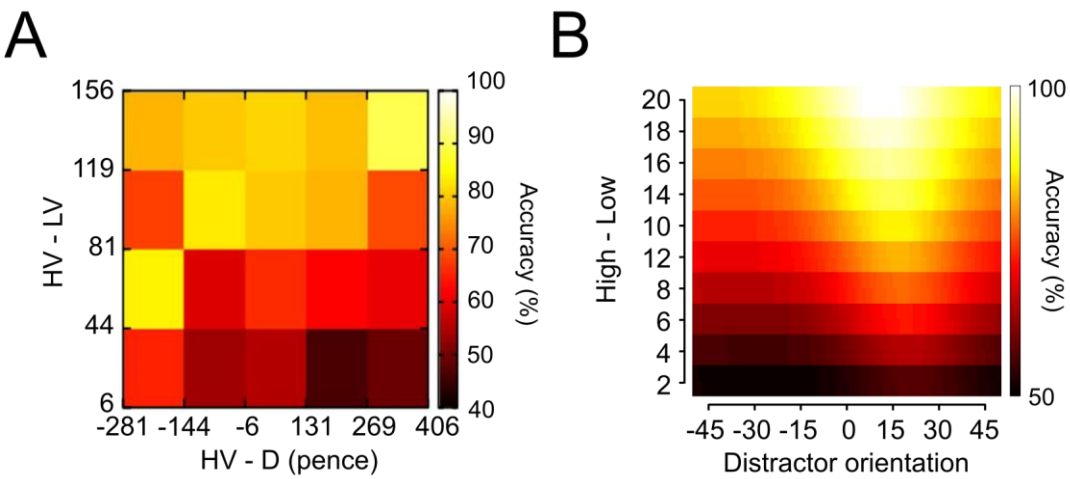

Figure S2. Simulated accuracy for a multi-alternative choice task. (A) Average choice accuracy in humans indicated by colour as a function of both the difference in value between the high valued and low valued options (HV − LV) and the difference in value between high valued options and the distractor (HV – D). Reprinted from Chau et al, permission pending. **(B)** Equivalent data from the adaptive gain model. The surface plot illustrates model accuracy (colour bar) for choosing the tilt that is further from the category boundary (‘High’) among tilts that are closer to the category boundary (‘Low’) as a function of a third distractor’s tilt. When the two tilts are dissimilar, the overall accuracy is higher than when the two tilts are similar. Accuracy is further modulated by the tilt of a third distractor that the accuracy is lower for a given dissimilarity between the two target tilts if the distractors are further away from the category boundary. Note similarities in the surface plot in (A).

Figure S3

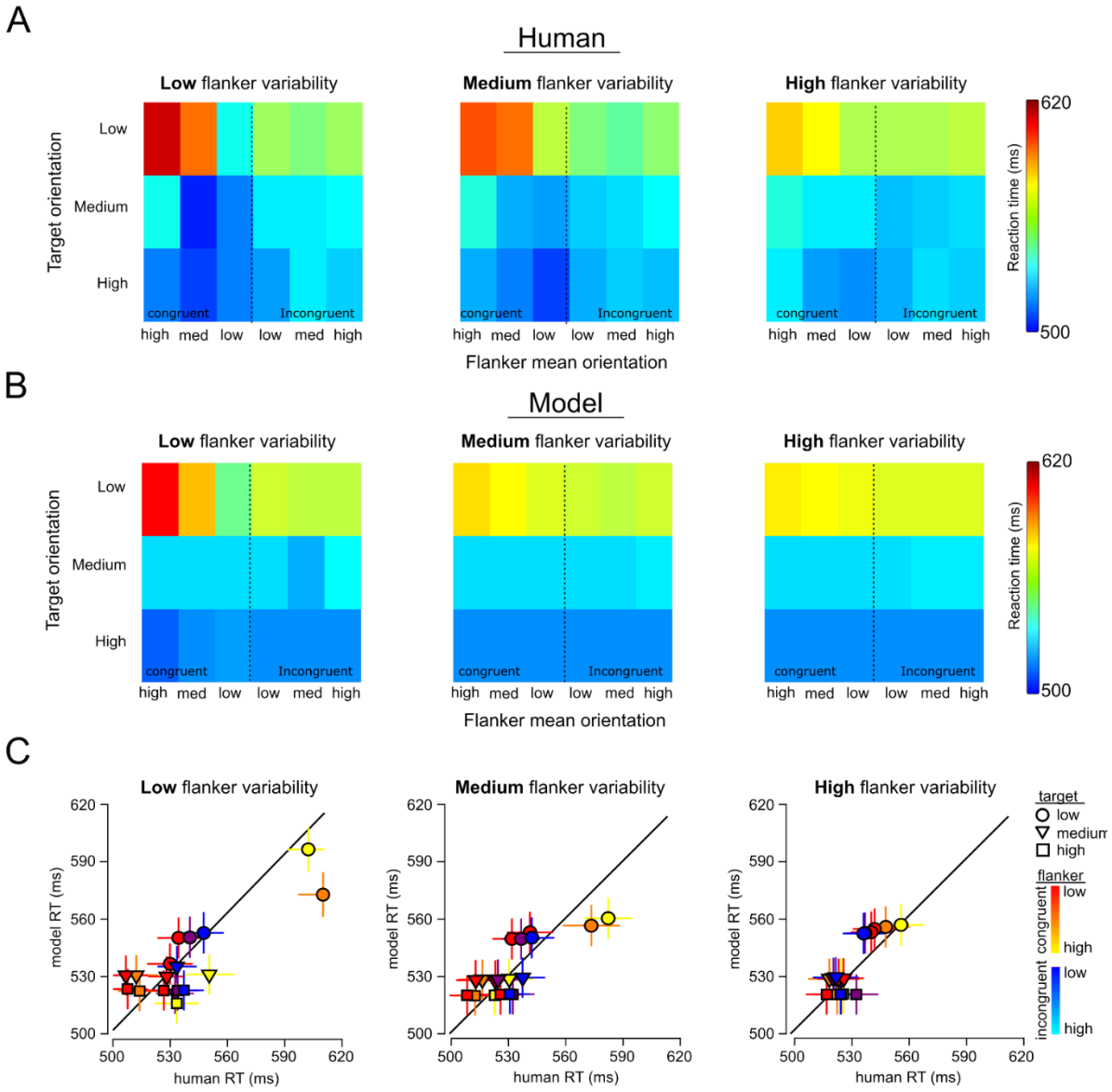

Figure S3. 4 way interactions for Exp. 2a (A) Model-predicted RTs for **Exp. 2a**; surface plot as for **Fig. 3A (B)** Human RTs for **Exp. 2a**; surface plot as for **Fig. 3B**, **(C)** Same cross-plot as **Fig. 3C**. See **Fig. 3** for details.

Figure S4

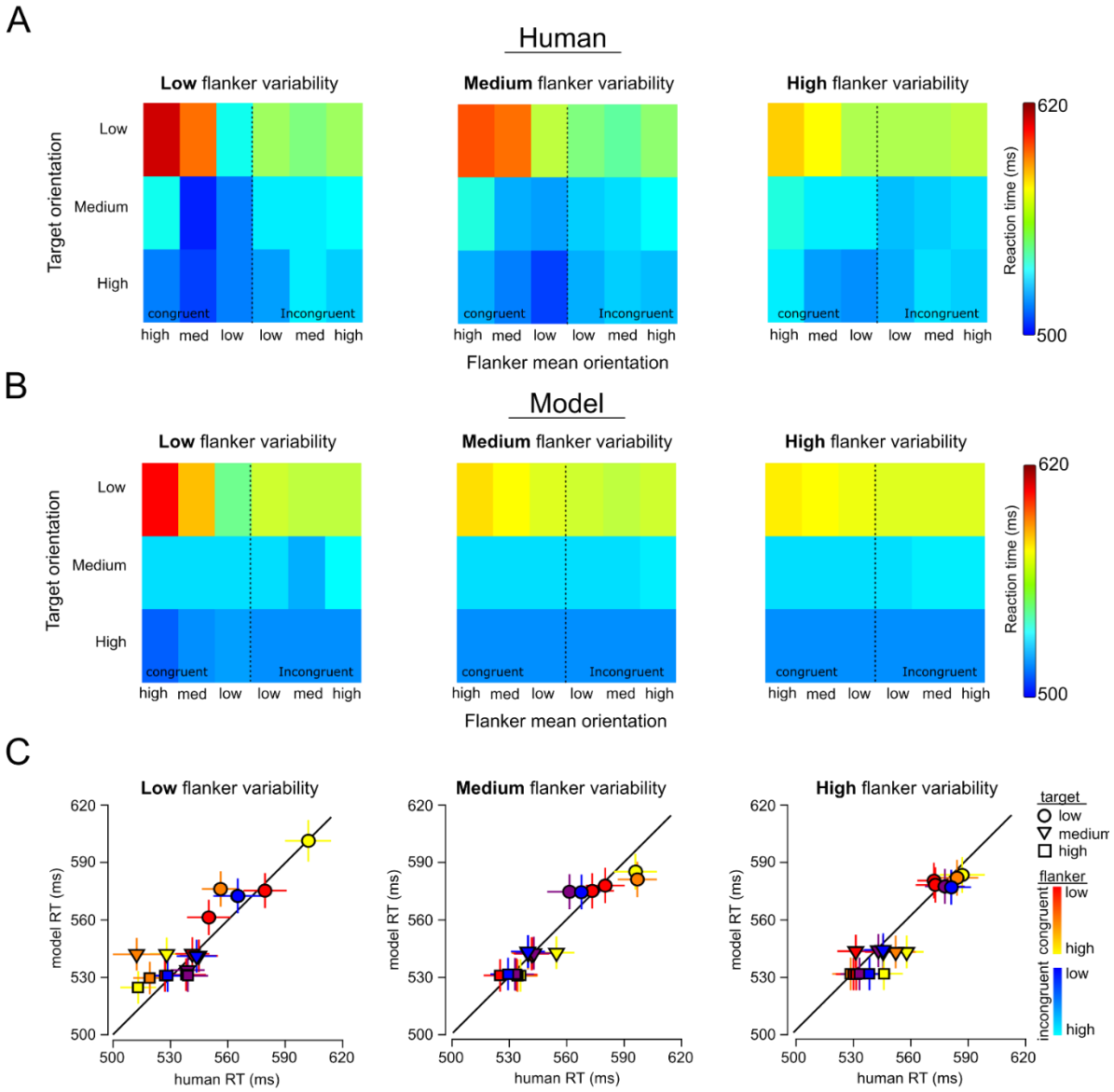

Figure S4. 4 way interactions for Exp. 2b (A, B) The same surface plots as **Fig. 3** and **Fig. S3 (C)** same cross-plots as **Fig. S2**. See Fig. S3 and Fig. 3 for details.

Fig. S5

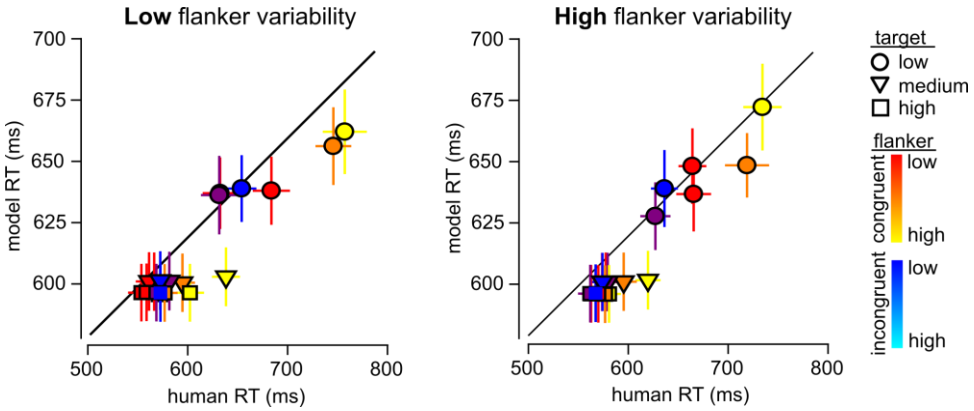

Figure S5. 4 way interactions for Exp.3. The same cross-plots as **Fig. S3c**

Table S1

Experiment	Target decision variable (X^i)	Flanker mean decision variable (mX^j)	Flanker standard deviation (sX^j)
1a	±45°	±45°	0, 15, 30
1b	±20°	±20°	5, 10, 15
2a	±15,30,45°	±15,30,45°	0,15,30
2b	±0.05, 0.1, 0.15	±0.05, 0.1, 0.15	0, 0.1, 0.2
3	U(-45,+45)	U(-45,+45)	U(0,30)

Table S1. Summary of experiments. Target decision variable, flanker mean decision variable and flanker variance were orthogonally manipulated in each experiment. Exp. 1 focused on the influence of flanker variance and congruence on RT. Exp. 2 introduces additional factors of target orientation and flanker mean in addition to flanker variance. In Exp. 3, the fMRI task, all variables were both parametric and orthogonal to each other. Values of each variable were drawn randomly from a uniform distribution with the range described in the methods.

Table S2

	p_{fdr}	Cluster extent	t	x	y	z
Target decision variable ($\lvert X^i \rvert$)	0.000	150	8.95	-2	-48	30
			7.11	-6	-52	14
			4.90	10	-48	38
	0.000	131	8.10	2	44	-14
			6.49	2	52	-10
			6.01	-2	32	-18
	0.000	83	7.92	-46	-60	30
			4.88	-38	-76	42
	0.000	62	6.68	-18	24	42
			5.25	-14	44	42
	0.001	45	5.86	54	-64	26
			5.57	46	-56	26
			5.40	62	-56	18
	0.004	31	5.14	58	-4	-22
			4.99	62	4	-22
			4.97	62	-12	-14
	0.017	20	4.72	42	16	-34
	0.000	36	9.65	14	-60	62
			8.19	22	-56	46
			8.12	34	-40	46
	0.000	14	8.43	-2	-24	-2
	0.000	12	8.31	54	-32	50
			7.89	46	-36	46
	0.000	29	8.23	34	24	2
			7.74	42	16	6
	0.000	15	8.21	-10	-12	6
	0.000	19	8.10	-42	-4	38
			7.86	-34	4	30
	0.000	18	7.98	-42	16	10
			7.94	-34	20	2
			7.71	-34	16	14

Table S2. Voxels correlating with $\lvert X^i \rvert$ at a threshold of $p < 0.05$ (FDR corrected) and with a cluster extent >10 voxels.

Table S3

	p_{fdr}	Cluster extent	t	x	y	z
Target decision variable modulated by the distance between target and flanker mean orientation ($\lvert X^i - mX^j \rvert \times \lvert X^i \rvert$)	0.000	195	8.23	26	-60	50
			7.69	30	-44	46
			7.39	14	-64	46
	0.000	164	7.18	30	20	10
			6.54	42	16	6
			6.28	46	8	14
	0.000	81	6.96	-30	20	10
			6.60	-38	12	-2
			4.63	-34	0	6
	0.000	110	6.14	-30	-44	54
			5.93	-26	-44	46
			4.52	-54	-28	38
	0.001	53	5.93	-38	4	30
			5.04	-50	0	38
	0.018	25	5.91	42	40	30
	0.030	21	5.89	30	-4	50
			4.81	22	-4	54
	0.001	55	5.36	6	16	46
			5.19	6	20	38
	0.001	50	5.20	-2	-20	-2
			5.08	10	-16	-10
			5.00	6	-24	-6
	0.015	27	4.89	-46	-60	-10
			4.71	-42	-60	2
	0.052	16	4.75	-10	-12	10
			4.22	-14	-4	6
	0.015	28	4.62	-10	-64	58
			4.06	-18	-60	66
			4.03	-14	-68	38
	0.019	30	4.60	-46	-72	34
			4.04	-46	-60	26
			3.78	-54	-64	30

Table S3. Voxels correlating with $|mX^j|$ at a threshold of $p < 0.05$ (FDR corrected) with a cluster extent > 10 voxels.

Table S4

	p_{fdr}	Cluster extent	t	x	y	z
Target decision variable modulated by the distance between target and flanker mean orientation ($\lvert X^i - mX^j \rvert \times \lvert X^i \rvert$)	0.000	195	8.23	26	-60	50
			7.69	30	-44	46
			7.39	14	-64	46
	0.000	164	7.18	30	20	10
			6.54	42	16	6
			6.28	46	8	14
	0.000	81	6.96	-30	20	10
			6.60	-38	12	-2
			4.63	-34	0	6
	0.000	110	6.14	-30	-44	54
			5.93	-26	-44	46
			4.52	-54	-28	38
	0.001	53	5.93	-38	4	30
			5.04	-50	0	38
	0.018	25	5.91	42	40	30
	0.030	21	5.89	30	-4	50
			4.81	22	-4	54
	0.001	55	5.36	6	16	46
			5.19	6	20	38
	0.001	50	5.20	-2	-20	-2
			5.08	10	-16	-10
			5.00	6	-24	-6
	0.015	27	4.89	-46	-60	-10
			4.71	-42	-60	2
	0.052	16	4.75	-10	-12	10
			4.22	-14	-4	6
	0.015	28	4.62	-10	-64	58
			4.06	-18	-60	66
			4.03	-14	-68	38
	0.019	30	4.60	-46	-72	34
			4.04	-46	-60	26
			3.78	-54	-64	30

Table S4. Voxels correlating with $\lvert X^i - mX^j \rvert \times \lvert X^i \rvert$ at a threshold of p < 0.05 (FDR corrected) with a cluster extent > 10 voxels.