Gain control explains the effect of distraction in human perceptual, cognitive, and economic decision making

Vickie Li[a,1], Elizabeth Michael[b], Jan Balaguer[a], Santiago Herce Castañón[a,c], and Christopher Summerfield[a]

[a]Department of Experimental Psychology, University of Oxford, OX2 6GG Oxford, United Kingdom; [b]Department of Psychology, University of Cambridge, CB2 3EB Cambridge, United Kingdom; and [c]Department of Psychology and Educational Sciences, University of Geneva, 1202 Geneva, Switzerland

Edited by Randolph Blake, Vanderbilt University, Nashville, TN, and approved July 3, 2018 (received for review March 26, 2018)

When making decisions, humans are often distracted by irrelevant information. Distraction has a different impact on perceptual, cognitive, and value-guided choices, giving rise to well-described behavioral 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 behavioral phenomena on variants of a classic cognitive paradigm, the Eriksen flanker task. We also report that blood oxygen level-dependent signals in a dorsal network prominently including the anterior cingulate cortex index a gain-modulated decision variable predicted by the model. This work unifies the study of distraction across perceptual, cognitive, and economic domains.

cognitive control | tilt illusion | decoy effects | gain control | anterior cingulate cortex

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 focusing exclusively on the choice-relevant information. This normative contention can be formalized in a variety of ways, for example via the notion that rational choices should be independent of irrelevant alternatives (1, 2) or that sensory signals should be weighted lawfully by their reliability and relevance to the choice at hand (3–6).

Nevertheless, empirical observations suggest that human decisions are unduly influenced by distracting information. Consider a generic problem in which a target stimulus X^i and distracters X^j occur at fixed spatial 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 categorized. 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. 1*A*) (7). 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) (8, 9). It is ubiquitously observed that incompatible flankers incur a cost, and compatible flankers confer a benefit, relative to a neutral condition, as measured in response times (RTs) and accuracy (Fig. 1*B*). Finally, behavioral and neural economists have charted the irrational influence that a decoy alternative of value Z has on choices between two choice-relevant prospects X and Y, where $X > Y$ (10–13). A common finding is that rational choices (i.e., for $X > Y$) initially decline as Z increases in value but then increase sharply as Z comes to approximately match the other two items in value (Fig. 1*C*); other stereotypical "decoy" effects are observed when alternatives X are characterized by more than one attribute (discussed below).

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 (14, 15), accounts that emphasize inhibitory interactions among competing sensory neurons or favor a normalization of stimulus values by a local average or range (10, 16–18), and Bayesian accounts that model spatial uncertainty among targets and distracters (19–21) or that assume a nonuniform prior on the compatibility of decision information (21). These accounts disagree about the computational mechanisms involved, the neural processing stages at which the cost of distraction arises, and the brain structures that are recruited to protect decisions against irrelevant information. For example, divisive normalization mechanisms may occur in sensory neurons in visual cortex (16), or among value representations in the orbitofrontal cortex (22), whereas the control systems that detect and resolve conflict have been attributed to medial and lateral prefrontal structures (14). 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 and shows that contextual signals determine the gain of processing of consistent (or "expected") features during decision-making

Significance

Information in the world can sometimes be irrelevant for our decisions. A good decision maker should take into account the relevant information and ignore the distracting information. However, empirical observation showed that human decisions are unduly influenced by distracting information. Diverse theories have been proposed to explain the cost that distracters incur during decision making across perceptual, cognitive, and economics domains. Here, we propose a single, unified model that is based on adaptive gain control to explain the influence of distraction across domains.

Author contributions: V.L., E.M., and C.S. designed research; V.L., J.B., and S.H.C. performed research; V.L., J.B., and C.S. contributed new reagents/analytic tools; V.L. and C.S. analyzed data; and V.L. and C.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: chui.li@psy.ox.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1805224115/-/DCSupplemental.

Published online August 30, 2018.

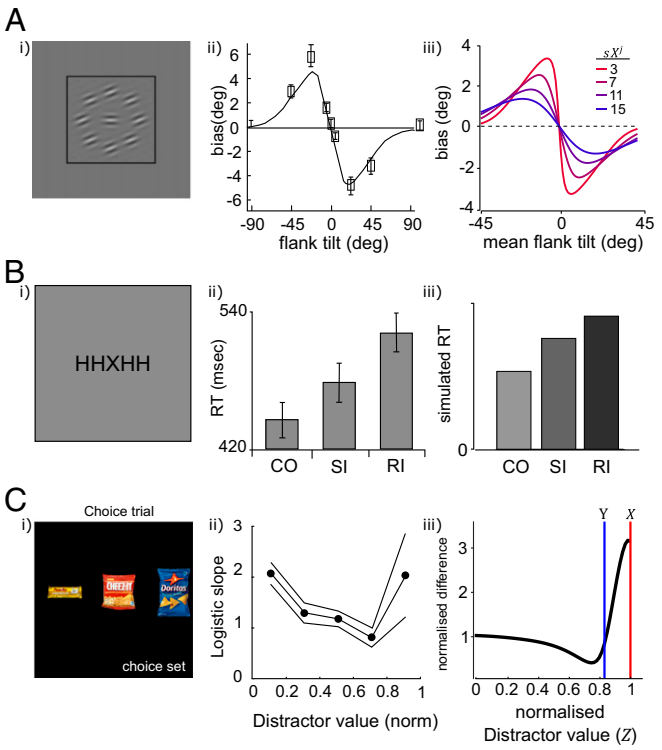

Fig. 1. The effect of distraction across perceptual, cognitive, and economic domains. (*A*, *i*) Participants were asked to discriminate the tilt (relative to horizontal) of a central Gabor surrounded by tilted distracters. (*A*, *ii*) Participants were biased to report the target as more clockwise when the flankers were counterclockwise, and vice versa (the "tilt illusion"). Panels *i* and *ii* republished with permission of Royal Society, from ref. 32; permission conveyed through Copyright Clearance Center, Inc. (*A*, *iii*) Simulation of the adaptive gain model replicates the qualitative tilt illusion pattern in ref. 32 and replicates the current human data both qualitatively and quantitatively (*SI Appendix*, Fig. S7). It further predicts that the magnitude of the bias is modulated by flanker variance; colored lines reflect flanker variabilities from low (red) to high (blue). (*B*, *i*) In the Eriksen flanker task, participants respond with a key press to a central letter while ignoring the flankers. (*B*, *ii*) RTs are the fastest on CO trials, then the SI trials, and slowest on RI trials. Panel *ii* republished with permission of MIT Press - Journals, from ref. 40; permission conveyed through Copyright Clearance Center, Inc. See *Methods* for details of how CO, RI, and SI trials were defined. (*B*, *iii*) The adaptive gain model predicts the same pattern of reaction time across the three conditions. (*C*, *i*) Participants chose the most preferred of three food items. (*C*, *ii*) Increasing the value of the least-preferred item reduces the choice efficiency (i.e., probability of choosing the highest-valued target) as the normalized distractor value increases, shown by logistic slope from fitting logistic choice functions on humans choice. Panel *ii* adapted with permission from ref. 10. (*C*, *iii*) Using the adaptive gain model, we simulated the normalized subjective difference between two options *X* and *Y* (blue and red line) as a function of a third decoy *Z* (*x* axis). The subjective difference is first reduced and the increased in a qualitatively similar fashion.

tasks (23). Our model, which is described here at the level of neural population codes, proposes that contextual signals sharpen the tuning curves of neurons with a compatible preference for decision-relevant features, and is motivated by a large literature emphasizing the need for adaptive gain control in the service of efficient coding (24, 25). 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 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 behavioral findings, including "reverse" compatibility effects (where fully visible, compatible flankers actually hinder, rather than help, behavioral performance). Over four behavioral experiments involving human participants, we validate these predictions, using visual stimuli defined by both tilt and color. Finally, we use functional brain imaging to show that the modulatory influence on decision signals predicted by the model correlates with blood oxygen level-dependent (BOLD) signals in the dorsal anterior cingulate cortex (dACC) and interconnected structures, where neural signals have variously been implicated in the context-sensitive encoding of action values (26), and the expected value of cognitive control (27). 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 (28).

Results

Our adaptive gain model is based on a framework that was previously developed to understand how humans performing spatial and temporal averaging tasks adapt to the context provided by proximal decision information (23, 29). Inputs arrive at a population of n decision neurons each characterized by a Gaussian tuning curve centered on its preferred feature value θ_k. Each neuron k responds to the target stimulus X^i with rate $R_k = f(X^i|\theta_k, \sigma_k)$, 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 linearly decoded into a subjective percept or value estimate $\widehat{X^i}$ by weighting the population activity R by the corresponding feature values θ:

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

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 (30), with a tuning width envelope that matches the inverse distribution of contextual features X^j with mean $\overline{X^j} = \sum_{j=1}^{6} X^j/d$ and standard deviation (SD) sX^j, where d is the number of distractors:

$$\sigma_k = \sigma^{max} - f\left(\theta_k|\overline{X^j}, sX^j + \varepsilon\right) \cdot n. \qquad [2]$$

In other words, neurons with a preferred orientation that matches $\overline{X^j}$ have the sharpest tuning curves, and these tuning curves are even sharper if the flanker variance (sX^j) is low (see Fig. 5*A*). In Eq. **2**, σ^{max} denotes the maximum tuning width in the population, ε is a constant parameter added to sX^j to ensure that the tuning widths are not zero, and n is a scaling parameter equivalent to the number of decision neurons.

We first show how the model explains both tilt illusion in perceptual choice tasks and decoy effects in economic choice tasks. In Fig. 1*A*, we plot the tilt bias over different values of $\overline{X^j} \in \{-45, -44 \ldots 45\}$ predicted by the model as a difference of subjective and objective estimates of the target $(\widehat{X^i} - X^i)$ for different $sX^j \in \{3, 7, 11, 15\}$. The model predicts that subjective estimates are repulsed away from the mean flanker value, as described in numerous previous studies (7, 31, 32), and additionally that the strongest repulsion effect occurs when the flankers are homogenous (i.e., they are drawn from a distribution with low dispersion). This repulsion of the subjectively decoded values from their objective counterparts occurs when the target feature is 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 *SI Appendix*, Fig. S1 for a more detailed explanation).

To model economic decoy effects, we envisage a choice between two prospects of value X and Y (where $X = 20$ and $Y = 10$ in arbitrary units, such as dollars) that is made in the context of distracters with a value Z. We plot the difference in their corresponding subjective estimates $\hat{X} - \hat{Y}$ as a function of Z, observing a pattern with a striking qualitative resemblance to that reported previously (10) (Fig. 1*C*). 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 Z increases, repulsion is first strongest toward Y (leading to a reduced preference for the objectively best option X) but then, as the decoy approaches the two items in the choice set, repulsion is maximal for X, reversing this effect. In further simulations, we systematically varied both the distance between X and Y, and Z, and we were able to capture the pattern of multialternative choice data described in a different study involving abstract shapes associated with different economic values (11) (*SI Appendix*, Fig. S2).

These simulations consider the influence of distractors on options that vary on a single decision-relevant dimension (e.g., tilt). However, decisions are often made about multiattribute stimuli, such as when a foraging animal evaluates fruit based on its color and size, or a consumer compares products with differing price and quality. A rich literature has shown that preferences for two otherwise equally preferred options can reverse in the presence of a third decoy stimulus, even when the decoy is less attractive or unavailable. For example, the choice between a powerful but more expensive laptop computer Y and a less powerful but more economical model X can be systematically biased toward X by the presence of a third option Z that is either yet more powerful and costly (compromise effect) (33), that is similar in power and price to Y (similarity effect) (13), or that is less powerful but more expensive than X (attraction effect) (34).

To test whether our model can also account for these choice biases, we computed subjective estimated values from the adaptive gain model independently for two attributes P and Q (with the gain field dictated by the decoy, Z) and summed them to provide a composite value estimate for the two alternatives: $\hat{X} = \hat{X}_P + \hat{X}_Q$; $\hat{Y} = \hat{Y}_P + \hat{Y}_Q$. We then plotted the relative preference $\hat{X} - \hat{Y}$ as a function of the position of Z in attribute space, yielding a surface plot that captured the attraction, compromise, and similarity effects (Fig. 2) in the manner described in various studies (17, 35, 36). Moreover, the model predicts that more extreme attraction and compromise decoys give rise to stronger effects, as previously described in a multiattribute gambling task (17).

Furthermore, our model predicts stereotyped patterns of intercorrelation among participants for the similarity, attraction, and compromise effects, with those who display a strong attraction effect also displaying a strong compromise effect but a weaker similarity effect (37, 38). To create plausible variation in sensitivity to context, we varied the tuning selectivities across 40 simulated participants and computed the correlation among each of the three decoy effects across the virtual cohort. As shown in Fig. 3, this analysis recreated the previously described interdependence in attraction, compromise, and similarity effects (37, 38). Finally, decoy effects have been shown to weaken when participants are placed under time pressure (39); by assuming that faster responses have lower signal-to-noise ratio at the decision formation stage, we can also recreate this feature of the data (Fig. 4 and *SI Appendix*, Fig. S3).

The Effect of Distractor Variance. Next, we used our model to simulate performance on a variant of the flanker task that involves categorizing a central grating X^i tilted at −45° from vertical, in the face of flanking gratings that are on average tilted in a compatible ($\overline{X^j} = -45°$) or incompatible ($\overline{X^j} = +45°$) fashion. In this setting, the model predicts slower RTs for incongurent trials, or for congruent but physically dissimilar flankers, as ubiquitously observed

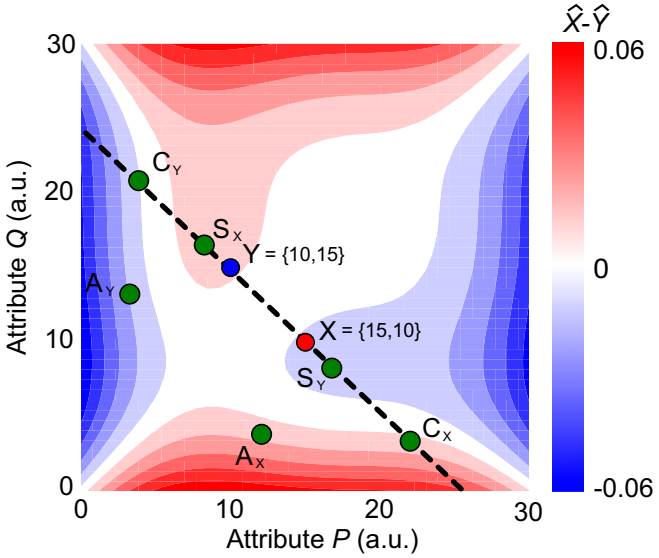

Fig. 2. The value difference between two equally preferred options (X and Y, red and blue dots) as a function of a third decoy option, Z (x and y axes). The axes correspond to the attribute values (e.g., inverse price [P] and quality [Q]). X and Y fall on the indifference line (dashed line). The red dot signals option X (where $X_P = 15, X_Q = 10$), and the blue dot indicates option Y (where $Y_P = 10, Y_Q = 15$). The colored surface shows the model-predicted subjective value difference $\hat{X} - \hat{Y}$ that is, the extent to which X is preferred over Y (red regions), and conversely the blue region in which the subjective value of X is smaller than Y, and thus Y is preferred. We have superimposed example decoy options (green dots) that produce the three context effects (compromise, C; attraction, A; and similarity, S; subscript corresponds to a preference for either option X or option Y). a.u., arbitrary units.

(40) (Fig. 1*B*). Because flanker effects for fully visible stimuli are strongest for RTs, we plot the inverse model output $1/|\widehat{X^i}|$ as a proxy for RT (i.e., we assume a ballistic evidence accumulation process with slope proportional to $|\widehat{X^i}|$; see Eq. **1** and *Methods*). Fig. 5*C* illustrates predictions from the adaptive gain model under two orthogonally varying factors: compatibility and flanker variability. 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, RTs should be reduced on compatible trials but remain the same on incompatible trials (Fig. 5*C*). 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 is similar at the tail of the gain distribution, and so the model predicts that flanker variance should not affect performance on incongruent trials (Fig. 5*A*). By contrast, classic models propose that response conflict varies with the amount of cross-talk interference among responses (14). These models predict that heightened flanker variance should have equal impact on compatible and incompatible trials (see *Methods* for model details).

Repulsion Effects in a Conflict Task. 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 six flanking distracter gratings (Fig. 5*B*; see task details in *SI Appendix*, *SI Materials and Methods*). The orientation of the target grating X^i and the mean of the flankers $\overline{X^j}$ were set to ±45° and the standard deviation of the flankers sX^j was varied at three levels: {0°, 15°, 30°} in Exp. 1a ($n = 37$; percent error = 5 ± 4.3 SD) and {5°, 10°, 15°} in Exp. 1b ($n = 36$; percent error = 4.9 ± 3.7 SD). Specifically, the zero flanker variance condition in Exp. 1a mimics the classic Eriksen flanker

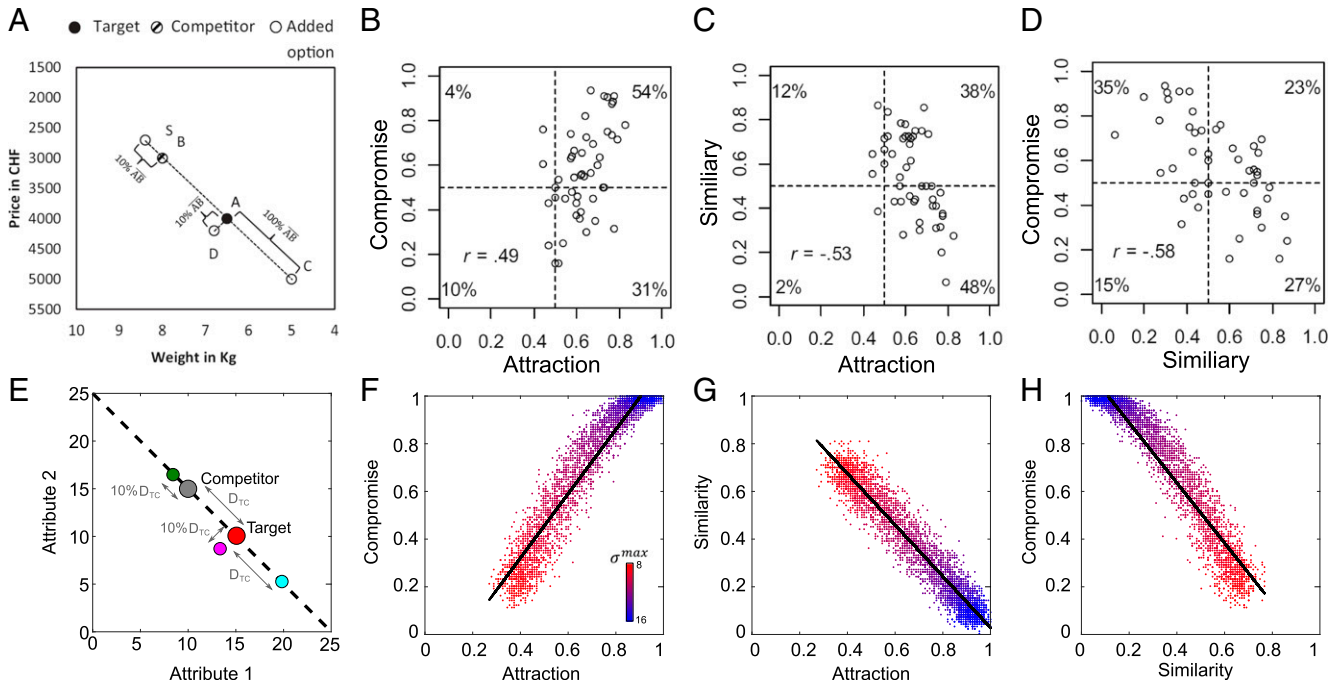

Fig. 3. Relative choice shares of the target for each pair decoy types. (*A–D*) Reprinted with permission from ref. 37. (*E–H*) Simulations from the adaptive gain model. (*E*) An illustration of the simulated 2D attribute space with target and decoys. The target (red dot) is set to be {15,10}, and the competitor (gray dot) is {10,15}. Both lie on the indifference line (dashed line). The three types of decoy and their values are plotted in three different colored circles (green: similarity decoys, value = {8.5,16.5}; magenta: attraction decoy, value = {13.5,8.5}; cyan: compromise decoys, value = {20,5}). D_{TC} refers to the distance between the target and the competitor. (*F–H*) Pairwise correlations among the strength of effect for each decoy type across the simulated cohort. Each unique colored dot corresponds to one simulated participant. The color corresponds to the tuning width σ^{max} (red = low tuning width σ^{max}, blue = high tuning width). The black line on each panel corresponds to the best least-squares fit. *F* shows a positive correlation between the compromise effect and the attraction effect across the population. *G* and *H* show a negative correlation between the similarity effect and the attraction or the compromise effect across the population. Across panels, we can see that participants who display a strong attraction effect also display a strong compromise effect but a weaker similarity effect, whereas those who display a weaker similarity effect display stronger compromise and attraction effects. This is consistent with the data displayed in *B–D* from ref. 37.

task, where flankers are identically tilted. In both experiments, we observed that flanker variance modulated RTs on compatible trials (Exp. 1a: $F_{1.93,69.45} = 10.11$, $P < 0.001$; Exp. 1b: $F_{2,67.91} = 9.66$, $P < 0.001$) but not on incompatible trials (both P values ≥ 0.25). This finding was qualified by an interaction between compatibility and flanker variance (Exp. 1a: $F_{1.76,63.39} = 9.72$, $P < 0.001$, Fig. 5*D*; Exp. 1b: $F_{1.96,66.79} = 9.67$, $P < 0.001$, Fig. 5*E*). Because previous work has shown that the ratio of compatible to incompatible flankers can modulate performance (41), 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 ($F_{1.72,61.9} = 10.61$, $P < 0.001$) and Exp. 1b ($F_{1.85,62.97} = 9.21$, $P < 0.001$). A full list of the ANOVA statistics and effect size on RT and accuracy for all experiments is reported in *SI Appendix*, Tables S2 and S3, respectively.

We fitted 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 (colored circles) are shown superimposed upon the human data in Fig. 5 *D* and *E*. 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 fits the human data more closely than the conflict model, with exceedance probabilities for the adaptive gain model of 0.9 in Exp. 1a and 0.58 in Exp. 1b (*SI Appendix*, Fig. S6*A*). We also compared a version of the gain model in which the contextual modulation was driven by both target and distracters; this model yielded both qualitatively and quantitatively similar results to the original gain model ($P > 0.3$ for both Exp. 1a and 1b), meaning that it is possible that contextual modulation arises from the entire array, rather than the flankers alone.

Next, we moved beyond the simple case in which X^i and $\overline{X^j}$ fell equidistant to the category boundary, using instead a more complex design where they could vary independently around vertical at {±15°, ±30°, ±45°}, and sX^j could once again vary at three levels {0°, 15°, 30°}. In this case, the model makes several predictions, some of them highly counterintuitive. First, 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 indistinguishable when the target (X^i) and flanker mean $\overline{X^j}$ are of the same or different sign. Second, 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 $\overline{X^j}$, for example, when $X^i = 15$ and $\overline{X^j} = 30$ or $\overline{X^j} = 45$ (Fig. 6*A*, upper left corner of each plot), and 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 $\overline{X^j}$ 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, $n = 28$; percent error = 3.76 ± 3.9 SD) performed the tilt categorization task described above, except with the full 6 (target mean) × 6 (flanker mean) × 3 (flanker variance) design. Another (Exp. 2b, $n = 30$; percent error = 8.47 ± 5.58 SD) performed a task with the same design that involved judging the color of a central circle (red vs. blue) surrounded by distracting flankers that varied continuously in color from red to blue. Results from the two experiments were qualitatively very similar (see *SI Appendix*, Figs. S4 and S5 for separate data) and so after normalizing the

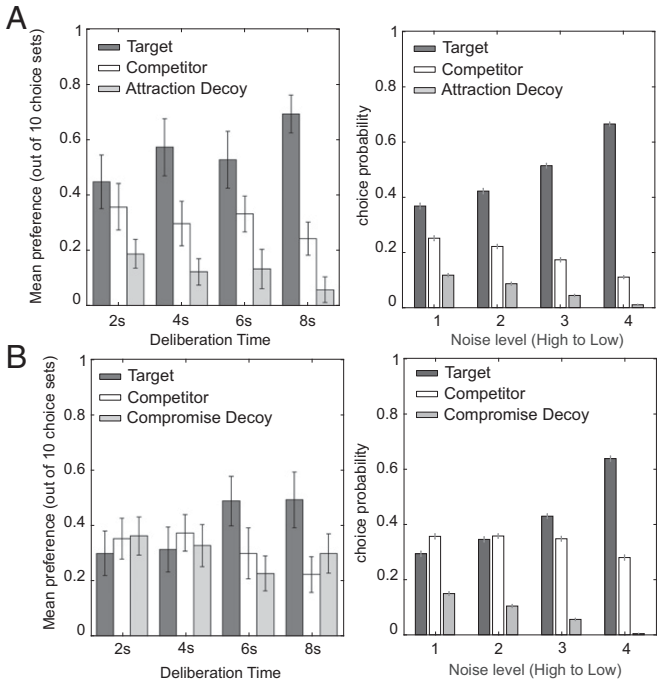

Fig. 4. Relative choice preference for the target, competitor, and decoy options under time pressure. (*A*) The target and competitor option is presented with an attraction decoy. (*B*) The target and competitor option is presented with a compromise decoy. (*Left*) Subjects' choice pattern under different deliberation time conditions. *A* and *B*, *Left* panels adapted with permission from ref. 39. (*Right*) Simulations from the adaptive gain model. Decreasing noise (comparable to increasing deliberation time) would increase the preference for the target option relative to the competitor option.

feature values (tilt and color) to an equivalent scale in the range [−180,180] we collapsed over them for display purposes. All three model predictions were strongly present in the human RT data (see Figs. 5*B* and 6 *A*–*C* for the fits of the conflict model). First, splitting trials naively into compatible and incompatible on this variant of the flanker task there was no overall significant difference in RT, either in the combined cohort ($P = 0.26$) or separately for each experiment (Exp. 2a: $P = 0.79$; Exp. 2b: $P = 0.17$). Second, 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 $X^i \times \overline{X^j}$ interaction (Exp. 2a: $F_{3.56,92.54} = 17.38$, $P < 0.001$; Exp. 2b: $F_{3.64,105.65} = 7.67$, $P < 0.001$), but not on incongruent trials (Exp. 2a: $P = 0.26$; Exp. 2b: $P = 0.23$). This difference was qualified by a reliable three-way $|X^i| \times |\overline{X^j}| \times$congruence interaction on human RTs (Exp. 2a: $F_{3.42,88.88} = 10.18$, $P < 0.001$; Exp. 2b: $F_{3.64,105.64} = 6.48$, $P < 0.001$). 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 individual flankers' tilt was similar or dissimilar to the target (42, 43). The adaptive gain model predicts a greater impact of flankers that are moderately dissimilar to targets compared with the case when they are highly dissimilar to the target (*SI Appendix*, Fig. S7), capturing the human data as well as replicating the reported influence of surround tilt on the perception of the central tilt from previous tilt illusion studies (32, 44).

Functional Brain Imaging. Established theories propose that a brain network that prominently includes the 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 on compatible trials (40, 45), and this effect is accentuated when the previous trial was compatible ("conflict adaptation"), as if the dACC is monitoring for conflict and signaling its onset (46). 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 (47), and its proximity to a choice point or category boundary (48); it responds to the relative economic value of an unchosen to a chosen option (49–51), to the value of switching to a new task or context (52, 53), and to the update signals that occur as decision values change (54). 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 (28, 55–57).

To assess the role of the dACC and interconnected regions in adaptive gain control, we conducted a new experiment in which X^i, $\overline{X^j}$, 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 while we acquired BOLD signals from across the brain using fMRI. Behavioral results of this experiment

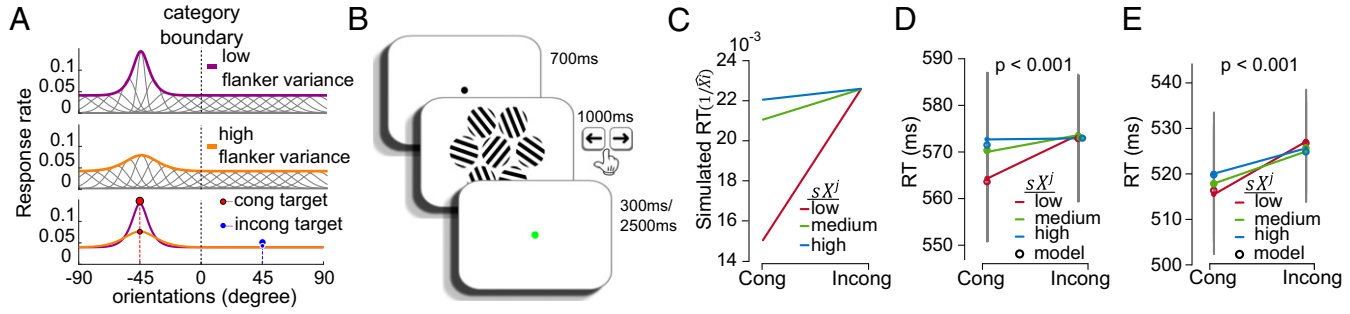

Fig. 5. Adaptive gain model of the Eriksen flanker task. (*A*) Illustration of the adaptive gain model as applied to Exp. 1. The model assumes that the population width tuning envelope is governed by the trial flanker mean ($\overline{X^j}$) and trial flanker variability (sX^j). Congruent targets (red 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 higher neural gain under low flanker variability trials (large circles) than under 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*) Task schematics. Participant first saw a fixation dot, followed by an array that contained a central target and six 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 ($1/|\overline{X^i}|$) using the adaptive gain model for three levels of flanker variance (sX^j, in colored lines) and congruency (*x* axis). (*D* and *E*) Human mean reaction time pattern (lines) under three levels of flanker variance and congruency condition in Exp. 1a (*D*) and Exp. 1b (*E*). Both interactions (congruency × flanker variance) are significant at $P < 0.001$. Colored circles represent the fitted mean reaction time from the adaptive gain model. Error bars show the SEM.

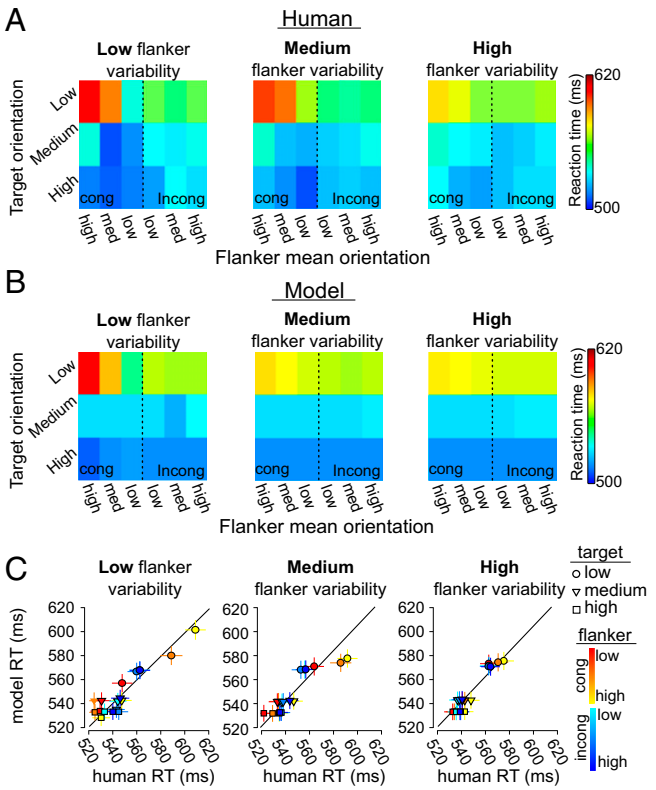

Fig. 6. Model predictions and human data for Exps. 2a and 2b. (*A*) Surface plots showing the mean RT pattern in humans under different conditions (three levels of target orientation × three levels of flanker mean orientation × congruency × three levels of flanker variability). Warmer colors correspond to longer RTs. There is an overall cost when the target is close to the category boundary (top row of each surface plot). There was an additional cost when these targets were flanked by congruent flankers that are further from the boundary (top left corner of each subplot). The introduction of higher flanker variability reduces these additional costs in those conditions (overall faster RTs across surface plots). (*B*) Fitted mean RT 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 colors (red to yellow) correspond to levels of mean orientation from congruent flankers. Cold colors (blue to cyan) correspond to levels of mean orientation from incongruent flankers $|\overline{X^j}|$. Different shapes correspond to three levels of target decision variable $|X^i|$.

replicated those from Exp. 2 (*SI Appendix*, Fig. S8), and so we focused 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, that is, in our experiment, when the target orientation is closer to vertical (48). 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, $t_{19} = 9.06$, $p_{fdr} < 0.001$; p_{fdr} denotes significance after correction for multiple comparisons using false discovery rate; *Methods*), as well as the anterior insula (AIC; peak: 34, 24, 2, $t_{19} = 10.19$, $p_{fdr} < 0.001$) and superior parietal lobe (SPL; peak: 22, −56, 46, $t_{19} = 8.19$, $p_{fdr} < 0.001$; see Fig. 7*A*). Extracting regions of interest from these areas in a leave-one-out fashion across participants (*Methods*), we then plotted how the BOLD signal varied in quartiles of both X^i and $\overline{X^j}$ (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. 7*B*). We found that the pattern of BOLD signals in all three regions closely resembled that predicted by the adaptive gain model, but not the other models (Fig. 7*C*). 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 $\overline{X^j}$ 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.

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 (*SI Appendix*, *SI Materials and Methods*) (58). 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 (*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 using random effects Bayesian model selection (59) and compared the resulting estimates to those obtained for rival models. Both the exceedance probabilities and the expected frequencies strongly favored the adaptive gain model over a decision model with no gain modulation, as well as over a family of conflict models (exceedance probabilities for the adaptive gain model in dACC: 0.992; AIC: 0.994; SPL: 0.996; expected frequencies for the adaptive gain model compared with chance level in dACC: $t_{19} = 4.11$, $P < 0.001$; AIC: $t_{19} = 4.11$, $P < 0.001$; SPL: $t_{19} = 4.76$, $P < 0.001$; Fig. 8*A*; see *SI Appendix*, *SI Materials and Methods* for the definition of the compared models). In other words, the dACC, along with AIC and SPL, codes for a

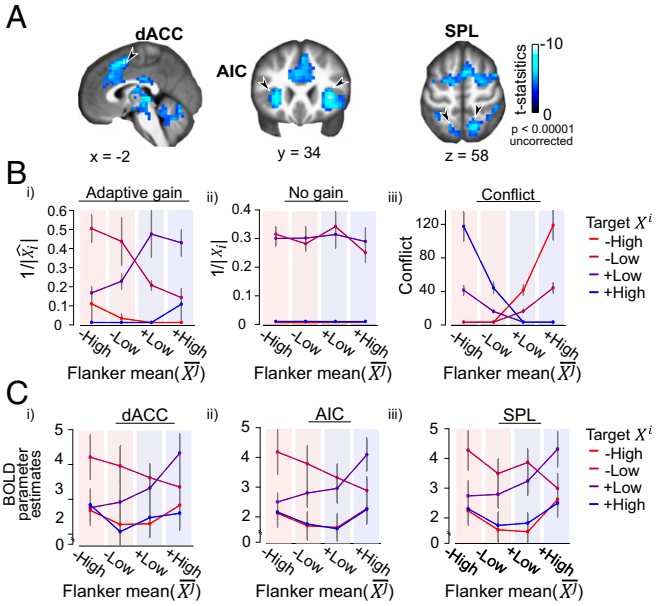

Fig. 7. Effect of target orientation on BOLD signal. (*A*) Brain areas correlating negatively with the absolute target decision variable, rendered onto a template brain in sagittal (*Left*), coronal (*Middle*), and axial (*Right*) slices. Images were generated with an uncorrected threshold of $P < 0.0001$. (*B*) Parametric modulators X^i and $\overline{X^j}$ were each split into four quartiles (lines for X^i, *x* axis and shaded area for $\overline{X^j}$). Mean predictions (±SEM) from different models: reciprocal gain-modulated decision variable $1/\hat{X}^i$ (*Left*). Reciprocal no-gain modulated target decision variable $1/\hat{X}^i$ (*Middle*). Estimated conflict was plotted for each quartile (*Right*). Colored lines correspond to four levels of X^i. Shaded colored background (groupings on *x* axis) corresponds to four levels of $\overline{X^j}$. (*C*) Mean BOLD (±SEM) signal beta values for each level of X^i and $\overline{X^j}$ from the quartile bins. (*Left*) dACC, (*Middle*) AIC, and (*Right*) SPL ROIs.

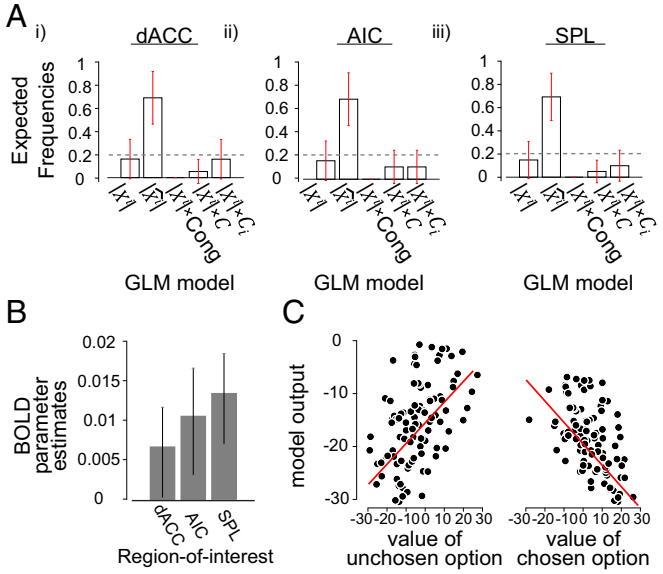

Fig. 8. (*A*) Model comparisons across GLM models. Expected frequencies for each GLM and each ROI were computed using random effects Bayesian model selection. Gray dotted line displayed the chance level see *SI Appendix, SI Materials and Methods* for the definition of the models. (*B*) ROI analysis on dACC, AIC, and SPL on the absolute flanker mean decision variable $|\overline{X^j}|$. Bar plots correspond to the mean (±SEM) beta estimates for this predictor from GLM4. (*C*) Simulated (negative) model output as a function of the value of chosen and unchosen option (see *Methods* for details).

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 (60). 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 (GLM3), we still recovered a significant activation in the dACC (peak: 6, 20, 50, t_{19} = 4.47, p_{fdr} = 0.046) and AIC (left peak: −30, 16, 10, t_{19} = 5.96, p_{fdr} = 0.038; right peak: 34, 24, 6, t_{19} = 5.05, p_{fdr} = 0.04; *SI Appendix*, Table S4). 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 (61), or to the relative value of an unchosen vs. a chosen option during economic choice (62)? We have already shown that in the simple version of the flanker task (cf. Exp. 1a where $sX^j = 0$), the model predicts a larger output $|\hat{X}_i|$ on compatible relative to incompatible 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 (63). 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 (64), just as it successfully accounts for the observation that during categorization of a multielement array RTs are faster if the target array is preceded by a prime array with an equivalent level of feature variance (29).

However, we also note another facet of our results: that BOLD signals in the dACC, AIC, and SPL regions of interest (ROIs) correlate negatively with $|X^j|$ but positively with $|\overline{X^j}|$ (GLM4; Fig. 8*B*; 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; *SI Appendix*, Table S5). 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 (*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 (65) (Fig. 8*C*). Our model thus unifies a number of disparate accounts that have emphasized a role for the dACC in tasks involving categorizing perceptual stimuli and choosing among economic prospects.

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 (1). 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 for single- and multiattribute stimuli. Here, we describe one such account. An adaptive gain model has previously been shown to successfully account for diverse contextual influences of relevant decision information in perceptual decision making, including confirmatory biases in sequential sampling (23) and priming by second-order summary statistics in perceptual categorization (29). Here, we show that not only can it account for classical effects of distraction across perceptual, cognitive, and economic domains, including the tilt illusion and decoy effects for both single- and multiattribute stimuli, it meets the elusive challenge of jointly capturing the attraction, compromise, and similarity effects, as well as their pattern of mutual dependence across participants (37, 38). Moreover, our model successfully predicts a range of previously unreported, counterintuitive findings in a well-studied cognitive paradigm, the Eriksen flanker task (8).

The effect of distraction is most often modeled 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 RTs and increases error rates (14). 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 mobilizing control mechanisms (associated with the lateral prefrontal cortex) to help mitigate the resulting costs (66). Here, we propose a differing view: one that emphasizes 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 efficiently for sensory inputs, reducing redundancy by dynamically adjusting the tuning properties of decision-relevant neurons to maximize sensitivity to expected features (24, 67). 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 (68), 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 et al. (64) more than 25 years ago.

We take the opportunity to highlight two major features of our behavioral data that cannot be accounted for by standard accounts that emphasize the cost of conflict alone. First, in Exps. 1a and 1b, we found that a low-variance flanker array hastens RTs on congruent trials, rather than prolonging RTs on incongruent trials. This is consistent with an account that emphasizes the benefit of consistent context rather than the cost of inconsistent context. Second, in Exps. 2a and 2b, we observed that the longest RTs were in fact observed on compatible trials, not incompatible trials. We replicated this finding across two different classes of visual feature: tilt and color. 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 reverse compatibility effects have been reported with heavily masked stimuli, where they can be explained by differing time courses of facilitatory vs. inhibitory processes (69), 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 (70). This report describes reverse compatibility effects when the target is most ambiguous, precisely parallelling our findings here for trials with small X^i and large $\overline{X^j}$, and a shift in the psychometric function that occurs with flanker identity in precisely the fashion predicted by our adaptive gain model (23).

Our behavioral 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 distracter's 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 does not ever incur an additional cost to accuracy and RTs, 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 alone fit our data more poorly.

Our findings present a challenge to some extant theories, but we acknowledge that our model is currently incomplete. For example, without further elaboration, our model cannot account for the previously described below-chance responding on the flanker task that is observed under strong speed pressure (71). Furthermore, a large literature implicates the dACC in the mechanisms by which we update the value of actions in dynamically changing environments (54, 72). 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.

Methods

Subjects. For behavioral studies Exps. 1 and 2, human participants were recruited via the online testing platform provided by Amazon Mechanical Turk (Exp. 1a: $n = 37$; Exp.1b: $n = 36$; Exp. 2a: $n = 28$; Exp. 2b: $n = 30$). For Exp. 3, 20 healthy volunteers with normal or corrected-to-normal vision and no history of neurological disorders were recruited to participate from the University of Granada, Spain. All participants gave informed consent to participate in the study and were compensated at a rate of $6 per hour for Exps. 1 and 2 and €10 per hour for the fMRI scanning session. All experiments were approved and conducted in accordance with the University of Oxford Medical Sciences Interdivisional Research Ethics Committee guidelines.

Design and Behavioral Analysis. For Exps. 1 and 2, the design orthogonalized the manipulation of target feature value (X^i), mean of flankers ($\overline{X^j}$), and variability of flankers (sX^j). We can further designate trials as "congruent" when X^i has the same sign as $\overline{X^j}$ or "incongruent" when X^i has the opposite sign as $\overline{X^j}$. In Exp. 1, we thus have a 3 × 2 (flanker variability × congruency) within-participant factorial design. In Exp. 2, we introduced three levels of $|\overline{X^j}|$ (equivalent to six levels of signed $\overline{X^j}$, three levels for each category), three levels of $|X^i|$, and three levels of sX^j, resulting in $3 \times 3 \times 3 \times 2$ ($|X^i| \times |\overline{X^j}| \times sX^j \times$ congruency) within-participant factorial design with 54 conditions. A full list of $|X^i|$, $|\overline{X^j}|$, and sX^j levels is displayed in *SI Appendix*, Table S1. For Exps. 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 behavioral analyses. Effect size – partial eta squared (η_p^2) was computed for all significant effects (*SI Appendix*, Tables S2 and S3). We only analyzed 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. These two exclusion criteria led to the following mean percentage (SD) of trials exclusion across subjects for each experiment: Exp. 1a, 6.85% (4.21); Exp. 1b, 6.73% (3.52); Exp. 2a, 5.62% (3.76); and Exp. 2b, 10% (5.44). We have also verified that all of the reported effects remained significant when we replaced mean RT with median RT or log-transformed RT.

Computational Modeling.

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 RT data (i.e., on a common scale in milliseconds), for each parameterization we regressed inverse decision values against each individual participant's RTs:

$$RT_{gain} = \beta^0 + \beta^1 \cdot 1/|\hat{X}_i|. \quad [3]$$

This calculation of RT is equivalent to modeling the data with ballistic (noiseless) diffusion process, with the two additional parameters β^0 and β^1 encoding the nondecision time and the drift rate, respectively (fixed across conditions). We used a balllistic accumulation process for simplicity, but note here that errors could be modeled by adding a noise term to the accumulation process. Searching exhaustively across values of σ^{max} and ε from Eq. **2**, we identified the parameters that minimized MSE between the human and model-predicted average RTs for each condition.

Conflict models. For the conflict model, we use a formulation described previously (14), whereby conflict C depends on the weighted product of competing inputs for the two actions A^{CW} and A^{CCW}:

$$C = \left[A^{CW} \cdot A^{CCW}\right] \quad [4]$$

$$A^{CW} = g\left(\overline{X^j}\right) \cdot (1 - w) + g(X^i) \cdot w$$

$$A^{CCW} = h\left(\overline{X^j}\right) \cdot (1 - w) + h(X^i) \cdot w,$$

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} are proportional to the tilt of the target and flanker mean. Alternatively, these values can also be defined according to the tilt of the target and each of the individual flankers X^j (conflict model 2):

$$C_i = \left[A^{CW}{}_i \cdot A^{CCW}{}_i\right] \quad [5]$$

$$A^{CW} = \sum_1^N g(X^j) \cdot \frac{(1 - w)}{d} + g(X^i) \cdot w$$

$$A^{CCW} = \sum_1^N h(X^j) \cdot \frac{(1 - w)}{d} + h(X^i) \cdot w,$$

where N is the number of flankers that are either congruent or incongruent, and d is the total number of flankers on a given trial. 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:

$$RT_{conflict} = \beta^0 + \beta^1 \cdot C. \quad [6]$$

To reduce the risk of overfitting we used cross-fitting, estimating model parameters on half the trials and computing MSE for the other half. These MSE were then fed into Bayesian model selection to compute exceedance probabilities (*SI Appendix, SI Materials and Methods*).

Simulations of Perceptual and Economic Decisions. When fitting to human data, the adaptive gain model contained two free parameters: maximum tuning width ($\sigma^{\max}$) and a constant parameter (ε). Simulations of the model aimed at qualitatively recreating effects from the past literature (e.g., for Fig. 1) assumed a fixed σ^{max} ($\sigma^{max} = 10$) and a fixed ε ($\varepsilon = 5$) unless noted otherwise. We have also imposed a floor value for σ_k (Eq. **2**). Any values that are below 0 will set as 0.1 so that the tuning width of any neurons will not be in negative values.

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 $\overline{X^j} \in \{-45, -44, \ldots, +45°\}$ and flanker SD $sX^j \in \{3,7,11,15\}$. For each variant of flanker mean decision values and flanker SD, $\widehat{X^i}$ is computed using Eqs. **1** and **2**. We then plot $\widehat{X^i}$ against levels of flanker mean decision value and flanker SD in Fig. 1*A*.

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

Decoy effect (single-attribute). We simulated the difference between the model estimated decision values from the two targets ($X = 20$ and $Y = 10$) as a function of a third distractor's decision value, $Z \in \{-45, -44, \ldots, +20\}$. ε was assumed to be 10 in this simulation. In *SI Appendix*, Fig. S2, we computed the model output associated with the highest-valued choice-relevant alternative ($\hat{X}$) and the next-best alternative ($\hat{Y}$), assuming that the mode of the gain field determined by the lowest-valued (i.e., irrelevant) alternative *Z*. We then plotted the normalized subjective estimates difference between the best option and the next-best option $(\hat{X} - \hat{Y})/\hat{Y}$, a quantity proportional to choice probability in Fig. 1*C*.

Decoy effects (multiattribute). We simulated the influence of a distracter (*Z*) on two equally preferred items *X* and *Y* that are characterized by two attributes, such as (inverse) price [*P*] and quality [*Q*]. We assume that the axes *P* and *Q* exhibit equal scaling and that that *X* and *Y* fall on the line of isopreference which lies perpendicular to the identity line. For illustration, we use $X = [15,10]$ and $Y = [10,15]$. We implemented the adaptive gain population coding model as follows:

$$\sigma_k = \sigma^{\max} - f(\theta_k | Z_P, \varepsilon) \cdot \tau. \quad [7]$$

Like in the single-attribute case, we assume that the inverse Gaussian tuning distribution is centered at the decoy attribute value (Z_P), with an SD $\varepsilon = 30$, and maximum tuning width $\sigma^{\max} = 10$ and an extra free parameter: tuning width range $\tau = 440$. Each neuron responds to the value of the attributes X_P and Y_P, resulting two hills of activity across the population of simulated neurons. The activity is decoded to subjective estimates for each of the two option values on attribute *P* ($\hat{X}_P$ and $\hat{Y}_P$) as described in Eq. **1** in the main text. The same procedure is repeated for attribute *Q*. Finally, the final value estimates for options *X* and *Y* are obtained by summation across attributes: $\hat{X} = \hat{X}_P + \hat{X}_Q$; $\hat{Y} = \hat{Y}_P + \hat{Y}_Q$. Fig. 2 shows the relative subjective value difference $(\hat{X} - \hat{Y})$ as a function of the the objective value of each attribute of a decoy *Z*, to produce a canonical 2D influence plot (heat map).

Intercorrelation between decoy types. We created a cohort of 40 simulated participants defined by differing levels of tuning selectivities (controlled by $\sigma^{\max}$; ranging from 8 to 16 with 40 linearly spaced values). Subjects with a high $\sigma^{\max}$ would have overall more broadly tuned tuning curves than subjects with a low σ^{max} (i.e., they are overall less susceptible to the effect of the context provided by the decoy). We then simulated three choice tasks (of 100 trials) each with a decoy type in the position shown for A, C, and S in Fig. 3*E* (i.e., decoys that have been shown to give rise to the attraction, compromise, and similarity effects in favor of the target). The exact values of these decoys were chosen in the same way as reported by Berkowitsch et al. (37) (Fig. 3*A*). We used the same two options, target (*X*) and competitor (*Y*), that are equally preferred when they are presented independently. Finally, the subjective estimated value difference between the target and the competitor is computed. To introduce intertrial variability in the simulations, we assumed that the value difference $(\hat{X} - \hat{Y})$ is corrupted by a noise term σ_n that is randomly sampled from a zero-mean Gaussian distribution with an SD of 0.08. The choice of a given trial is simply based on the sign of the noisy value difference between $\hat{X}$ and $\hat{Y}$. Averaging across trials allows us to compute the probability of choosing the target. We repeated the same procedure 100 times to obtain a more accurate estimation of the choice pattern. The choice probabilities from each decoy task were plotted against each other to visualize the correlation and anticorrelation of decoy effects across subjects (Fig. 3 *F–H*).

The effect of noise at decision-formation stage on decoy effects. In Fig. 4, we simulated the attraction effect with the following options and parameters: $X = \{15,10\}$; $Y = \{10,15\}$; $Z = \{13,5\}$; $\sigma^{max} = 20$; $\varepsilon = 10$; and $\tau = 400$. After obtaining decoded subjective estimates $\hat{X}$, $\hat{Y}$, and $\hat{Z}$, we then add noise to these subjective estimates by sampling from zero-mean Gaussian distribution with differing SDs (σ_n; four linearly spaced values between 3–7). The simulated choice on a given trial was based on the option with the highest noisy subjective estimates among options *X*, *Y*, and *Z*. Choice probability is computed by the proportion of trials one selected for each option. We repeated the process 100 times to obtain accurate estimates of the choice probability. We carried out the same analysis with a compromise decoy = {20,5}. For this analysis, we used a slightly different set of parameters: $\sigma^{max} = 40$, $\tau = 140$, and $\sigma_n \in [0.08 \ldots 0.3]$. These parameters allow us to recreate most faithfully the compromise effect choice pattern under time pressure shown in the Pettibone study (39), but similar effects were obtained for simulations with different parameters within a reasonable range. The effect of noise on similarity decoys under inferential and perceptual tasks like in Trueblood et al. (73) is reported in *SI Appendix*, Fig. S3.

Value of chosen vs. unchosen option. We simulated the model output as a function of the value of a theoretical chosen and unchosen option with an $\varepsilon = 10$. On each trial, decision values for two stimuli (X^1 and X^2) were drawn independently from two zero-mean Gaussian distribution with an SD 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 a target with the alternative as a 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:

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

$$CP = \frac{1}{1 + e^{\frac{-D}{5}}}. \quad [8]$$

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

fMRI Data Analyses. We analyzed our data using statistical parametric mapping (SPM12) with the general linear model (GLM) framework and in-house scripts running in MATLAB. For all analyses, we ensured that sequential orthogonalization of predictors in SPM was disabled. All GLMs also included regressors encoding the estimated movement parameters from preprocessing as a nuisance covariate. We modeled trials by convolving regressors coding the onsets and durations of events with the canonical hemodynamic response function and regressed them against the BOLD signal. Error trials, and those for which RT fell within the most extreme percentiles (<1% or >99%), were modeled separately as a nuisance regressor in all GLMs (trial exclusion: 6.68% ± 3.44 SD). 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 that correlated negatively with this regressor to define ROIs in the dACC, AIC, and SPL. Activations in these regions survived false discovery rate (fdr) correction for multiple comparisons at $p_{fdr} < 0.05$. To avoid double dipping, 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 an ROI with a threshold of uncorrected $P < 0.0001$, from which beta values were extracted from the left-out participant. For GLM2, we discretize each parametric modulator (*i*) target decision values X^i and (*ii*) flanker mean feature value $\overline{X^j}$ into four quartiles bins and included a total of $4 \times 4 = 16$ regressors (corresponding to each quartile bins of feature values) in the GLM. BOLD betas from each subject were again extracted using each ROI mask defined by the leave-one-out analysis. In GLM3, we included empirically observed RT as a competing regressor to $1/|\hat{X}_i|$ (a full list of active voxels for $1/|\hat{X}_i|$ are displayed in *SI Appendix*, Table S4). In GLM4, 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 $|\overline{X^j}|$, (iii) flanker precision (i.e., inverse flanker variability $1/sX^j$), (iv) absolute distance between target and flanker mean orientations $|X^i - \overline{X^j}|$, (v) the interaction of the distance between target and flanker mean with target decision values $|X^i - \overline{X^j}| \times |X^i|$, and ($vi$) the interaction of the distance between target and flanker mean, target decision values, and flanker precision ($|X^i - \overline{X^j}| \times |\overline{X^j}| \times 1/sX^j$). Full details of active voxels associated with these regressors are displayed in *SI Appendix*, Table S5.

Subsequent statistical tests on the betas extracted from each ROI were conducted using a two-tailed one-sample t test at a significance threshold of $P < 0.05$.

ACKNOWLEDGMENTS. We thank Tobias Egner and two anonymous reviewers for comments on an earlier version of this manuscript. This work was funded by European Research Council Starter Grants 281628 and 725973 (to C.S.), a Croucher Foundation scholarship (to V.L.), a Medical Research Council studentship (to E.M.), and Wellcome Trust Grant 0099741/Z/12/Z (to S.H.C.).

1. Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis* (Wiley, New York).
2. von Neumann J, Morgenstern O (1944) *Theory of Games and Economic Behavior* (Princeton Univ Press, Princeton).
3. Savage LJ (1954) *The Foundations of Statistics* (Wiley, New York).
4. Wald A, Wolfowitz J (1949) Bayes solutions of sequential decision problems. *Proc Natl Acad Sci USA* 35:99–102.
5. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. *Nat Neurosci* 9:1432–1438.
6. Körding K (2007) Decision theory: What "should" the nervous system do? *Science* 318:606–610.
7. Blakemore C, Carpenter RH, Georgeson MA (1970) Lateral inhibition between orientation detectors in the human visual system. *Nature* 228:37–39.
8. Eriksen BA, Eriksen CW (1974) Effects of noise letters upon identification of a target letter in a non- search task. *Percept Psychophys* 16:143–149.
9. Kopp B, Mattler U, Rist F (1994) Selective attention and response competition in schizophrenic patients. *Psychiatry Res* 53:129–139.
10. Louie K, Khaw MW, Glimcher PW (2013) Normalization is a general neural mechanism for context-dependent decision making. *Proc Natl Acad Sci USA* 110:6139–6144.
11. Chau BK, Kolling N, Hunt LT, Walton ME, Rushworth MF (2014) A neural mechanism underlying failure of optimal choice with multiple alternatives. *Nat Neurosci* 17:463–470.
12. Tsetsos K, Usher M, Chater N (2010) Preference reversal in multiattribute choice. *Psychol Rev* 117:1275–1293.
13. Tversky A (1972) Elimination by aspects: A theory of choice. *Psychol Rev* 79:281–299.
14. Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. *Psychol Rev* 108:624–652.
15. Bundesen C (1998) A computational theory of visual attention. *Philos Trans R Soc Lond B Biol Sci* 353:1271–1281.
16. Heeger DJ (1992) Normalization of cell responses in cat striate cortex. *Vis Neurosci* 9:181–197.
17. Soltani A, De Martino B, Camerer C (2012) A range-normalization model of context-dependent choice: A new model and evidence. *PLOS Comput Biol* 8:e1002607.
18. Schwartz O, Sejnowski TJ, Dayan P (2009) Perceptual organization in the tilt illusion. *J Vis* 9:19.1–19.20.
19. Dayan P, Solomon JA (2010) Selective Bayes: Attentional load and crowding. *Vision Res* 50:2248–2260.
20. Logan GD (1996) The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychol Rev* 103:603–649.
21. Yu AJ, Dayan P, Cohen JD (2009) Dynamics of attentional selection under conflict: Toward a rational Bayesian account. *J Exp Psychol Hum Percept Perform* 35:700–717.
22. Padoa-Schioppa C (2009) Range-adapting representation of economic value in the orbitofrontal cortex. *J Neurosci* 29:14004–14014.
23. Cheadle S, et al. (2014) Adaptive gain control during human perceptual choice. *Neuron* 81:1429–1441.
24. Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. *Annu Rev Neurosci* 24:1193–1216.
25. Rangel A, Clithero JA (2012) Value normalization in decision making: Theory and evidence. *Curr Opin Neurobiol* 22:970–981.
26. Kolling N, et al. (2016) Value, search, persistence and model updating in anterior cingulate cortex. *Nat Neurosci* 19:1280–1285.
27. Shenhav A, Cohen JD, Botvinick MM (2016) Dorsal anterior cingulate cortex and the value of control. *Nat Neurosci* 19:1286–1291.
28. Heilbronner SR, Hayden BY (2016) Dorsal anterior cingulate cortex: A bottom-up view. *Annu Rev Neurosci* 39:149–170.
29. Michael E, de Gardelle V, Summerfield C (2014) Priming by the variability of visual information. *Proc Natl Acad Sci USA* 111:7873–7878.
30. Gilbert CD, Wiesel TN (1990) The influence of contextual stimuli on the orientation selectivity of cells in primary visual cortex of the cat. *Vision Res* 30:1689–1701.
31. Gibson JJ (1937) Adaptation with negative aftereffect. *Psychol Rev* 44:222–244.
32. Solomon JA, Morgan MJ (2006) Stochastic re-calibration: Contextual effects on perceived tilt. *Proc Biol Sci* 273:2681–2686.
33. Simonson I (1989) Choice based on reasons: The case of attraction and compromise effects. *J Consum Res* 16:158–174.
34. Huber J, Payne JW, Puto C (1982) Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *J Consum Res* 9:90–98.
35. Rigoli F, Mathys C, Friston KJ, Dolan RJ (2017) A unifying Bayesian account of contextual effects in value-based choice. *PLoS Comput Biol* 13:e1005769.
36. Landry P, Webb R (May 14, 2018) Pairwise normalization: A neuroeconomic theory of multi-attribute choice. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2963863.
37. Berkowitsch NA, Scheibehenne B, Rieskamp J (2014) Rigorously testing multialternative decision field theory against random utility models. *J Exp Psychol Gen* 143:1331–1348.
38. Liew SX, Howe PD, Little DR (2016) The appropriacy of averaging in the study of context effects. *Psychon Bull Rev* 23:1639–1646.
39. Pettibone JC (2012) Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice. *Judgm Decis Mak* 7:513–523.
40. Van Veen V, Carter CS (2002) The timing of action-monitoring processes in the anterior cingulate cortex. *J Cogn Neurosci* 14:593–602.
41. Wu T, Dufford AJ, Mackie MA, Egan LJ, Fan J (2016) The capacity of cognitive control estimated from a perceptual decision making task. *Sci Rep* 6:34025.
42. de Gardelle V, Summerfield C (2011) Robust averaging during perceptual judgment. *Proc Natl Acad Sci USA* 108:13341–13346.
43. Li V, Herce Castañón S, Solomon JA, Vandormael H, Summerfield C (2017) Robust averaging protects decisions from noise in neural computations. *PLoS Comput Biol* 13:e1005723.
44. Solomon JA, Morgan MJ (2009) Strong tilt illusions always reduce orientation acuity. *Vision Res* 49:819–824.
45. Carter CS, et al. (1998) Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* 280:747–749.
46. Kerns JG, et al. (2004) Anterior cingulate conflict monitoring and adjustments in control. *Science* 303:1023–1026.
47. Ho TC, Brown S, Serences JT (2009) Domain general mechanisms of perceptual decision making in human cortex. *J Neurosci* 29:8675–8687.
48. Grinband J, Hirsch J, Ferrera VP (2006) A neural representation of categorization uncertainty in the human brain. *Neuron* 49:757–763.
49. Wunderlich K, Rangel A, O'Doherty JP (2009) Neural computations underlying action-based decision making in the human brain. *Proc Natl Acad Sci USA* 106:17199–17204.
50. Tsetsos K, Wyart V, Shorkey SP, Summerfield C (2014) Neural mechanisms of economic commitment in the human medial prefrontal cortex. *eLife* 3:e03701.
51. Juechems K, Balaguer J, Ruz M, Summerfield C (2017) Ventromedial prefrontal cortex encodes a latent estimate of cumulative reward. *Neuron* 93:705–714.e4.
52. Kolling N, Behrens TE, Mars RB, Rushworth MF (2012) Neural mechanisms of foraging. *Science* 336:95–98.
53. Barber AD, Carter CS (2005) Cognitive control involved in overcoming prepotent response tendencies and switching between tasks. *Cereb Cortex* 15:899–912.
54. Rushworth MF, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal cortex and reward-guided learning and decision-making. *Neuron* 70:1054–1069.
55. Shenhav A, Botvinick MM, Cohen JD (2013) The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron* 79:217–240.
56. Kolling N, Behrens T, Wittmann MK, Rushworth M (2016) Multiple signals in anterior cingulate cortex. *Curr Opin Neurobiol* 37:36–43.
57. Shahnazian D, Holroyd CB (2018) Distributed representations of action sequences in anterior cingulate cortex: A recurrent neural network approach. *Psychon Bull Rev* 25:302321.
58. Penny W, Mattout J, Trujillo-Barreto N (2006) Bayesian model selection and averaging. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, eds Friston K, Ashburner J, Kiebel S, Nichols T, Penny W (Elsevier, London).
59. Daunizeau J, Adam V, Rigoux L (2014) VBA: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput Biol* 10:e1003441.
60. Grinband J, et al. (2011) The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. *Neuroimage* 57:303–311.
61. Botvinick M, Nystrom LE, Fissell K, Carter CS, Cohen JD (1999) Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature* 402:179–181.
62. Boorman ED, Rushworth MF, Behrens TE (2013) Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. *J Neurosci* 33:2242–2253.
63. Hazeltine E, Poldrack R, Gabrieli JD (2000) Neural activation during response competition. *J Cogn Neurosci* 12:118–129.
64. Gratton G, Coles MG, Donchin E (1992) Optimizing the use of information: Strategic control of activation of responses. *J Exp Psychol Gen* 121:480–506.
65. Boorman ED, Behrens TE, Woolrich MW, Rushworth MF (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.
66. Egner T, Hirsch J (2005) Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nat Neurosci* 8:1784–1790.
67. Barlow H (1961) Possible principles underlying the transformation of sensory messages. *Sensory Communication* (MIT Press, Cambridge, MA).
68. Stemmler M, Usher M, Niebur E (1995) Lateral interactions in primary visual cortex: A model bridging physiology and psychophysics. *Science* 269:1877–1880.
69. Eimer M (1999) Facilitatory and inhibitory effects of masked prime stimuli on motor activation and behavioural performance. *Acta Psychol (Amst)* 101:293–313.
70. Rouder JN, King JW (2003) Flanker and negative flanker effects in letter identification. *Percept Psychophys* 65:287–297.
71. Gratton G, Coles MG, Sirevaag EJ, Eriksen CW, Donchin E (1988) Pre- and poststimulus activation of response channels: A psychophysiological analysis. *J Exp Psychol Hum Percept Perform* 14:331–344.
72. Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10:1214–1221.
73. Trueblood JS, Brown SD, Heathcote A (2014) The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychol Rev* 121:179–205.