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

Vickie Li1,a, Elizabeth Michaelb, Jan Balaguerb, Santiago Herce CastañoAc,d, and Christopher Summerfielda

1Department of Experimental Psychology, University of Oxford, OX2 6GG Oxford, United Kingdom; bDepartment of Psychology, University of Cambridge, CB2 3EJ Cambridge, United Kingdom; and cDepartment 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 counter-intuitive 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′ and distracters Xoccur 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′ (e.g., a central grating) is repulsed away from the mean tilt of X (surrounding gratings with similar but nonidentical tilt; Fig. L1)(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 based on adaptive gain control to explain the influence of distraction across domains.

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 economic domains. Here, we propose a single, unified model that is based on adaptive gain control to explain the influence of distraction across domains.

By implementing the idea that contextual signals determine the gain of processing mechanisms, the proposed model accommodates a rich count of its implementation across count of its implementation across perceptual, cognitive, and economic domains. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

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.

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 economic 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.

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 economic domains. Here, we propose a single, unified model that is based on adaptive gain control to explain the influence of distraction across domains.
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 flanks 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 flanks. (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 human choices. 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.

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 flanks 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 flanks 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 θk. Each neuron k responds to the target stimulus X′ with rate $R_k = f(X'/\theta_k, \sigma_k)$, where $f(X'/\theta, \sigma)$ denotes the probability density function of the normal distribution with mean θ and variance $\sigma^2$.

The estimated output of the neural population is then linearly decoded into a subjective percept or value estimate $\hat{X'}$ by weighing the population activity by the corresponding feature values $\theta_j$.

$$\hat{X'} = \sum_{k=1}^{n} R_k \cdot \theta_k , \quad [1]$$

When the gain is uniformly spread across the feature space (i.e., the tuning widths $\sigma_k$ 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 distractors modulates the sharpening of neuronal tuning (30), with a tuning width envelope that matches the inverse distribution of contextual features $X'$ with mean $\hat{X'} = \sum_{j=1}^{n} X'/d$ and standard deviation (SD) $sX'$, where $d$ is the number of distractors:

$${\sigma}_k = s_{\text{max}} \cdot f(\theta_k; \hat{X'}, sX' + \epsilon) \cdot n. \quad [2]$$

In other words, neurons with a preferred orientation that matches $\hat{X'}$ have the sharpest tuning curves, and these tuning curves are even sharper if the flanker variance ($sX'$) is low (see Fig. 5A). In Eq. 2, $s_{\text{max}}$ denotes the maximum tuning width in the population, $\epsilon$ is a constant parameter added to $sX'$ 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. 1A, we plot the tilt bias over different values of $\hat{X'} \in (-45, -44...45)$ predicted by the model as a difference of subjective and objective estimates of the target ($\hat{X'} - X'$) for different $sX' \in \{3, 7, 11, 15\}$. 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 flanks are homogenous (i.e., they are drawn from a distribution with low dispersion). This repulsion of the subjective 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 $t_i$ (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 distractors with a value $Z$. We plot the difference in their corresponding subjective estimates $X - Y$ as a function of $Z$, observing a pattern with a striking qualitative resemblance to that reported previously (10) (Fig. 1C). 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 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: $\tilde{X} = X_P + X_Q$; $\tilde{Y} = Y_P + Y_Q$). We then plotted the relative preference $X - Y$ 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 $X^c$ tilted at $-45^\circ$ from vertical, in the face of flanking gratings that are on average tilted in a compatible ($\tilde{X} = -45^\circ$) or incompatible ($\tilde{X} = +45^\circ$) fashion. In this setting, the model predicts slower RTs for incongruent trials, or for congruent but physically dissimilar flankers, as ubiquitously observed (40) (Fig. 1B). Because flanker effects for fully visible stimuli are strongest for RTs, we plot the inverse model output $1/|X^c|$ as a proxy for RT (i.e., we assume a ballistic evidence accumulation process with slope proportional to $|X^c|$; see Eq. 1 and Methods). Fig. 5C 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 distractors were of congruent sign. However, it also makes a new, testable prediction: that as $X^c$ (flanker variance) decreases, RTs should be reduced on compatible trials but remain the same on incompatible trials (Fig. 5C). This occurs because on compatible trials (flankers at $-45^\circ$), 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 $+45^\circ$) is negligibly different across different flanker variance levels since the neural gain they receive 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. 5A). 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 distractor gratings (Fig. 5B; see task details in SI Appendix, SI Materials and Methods). The orientation of the target grating $X^c$ and the mean of the flankers $\tilde{X}$ were set to $\pm 45^\circ$ and the standard deviation of the flankers $sX^c$ was varied at three levels: $(0^\circ, 15^\circ, 30^\circ)$ in Exp. 1a ($n = 37$; percent error $= 5 \pm 4.3$ SD) and $(5^\circ, 10^\circ, 15^\circ)$ in Exp. 1b ($n = 36$; percent error $= 4.9 \pm 3.7$ SD). Specifically, the zero flanker variance condition in Exp. 1a mimics the classic Eriksen flanker
task, where flankers are identically tilted. In both experiments, we observed that flanker variance modulated RTs on compatible trials (Exp. 1a: $F_{9,3,64,91} = 10.11, P < 0.001$; Exp. 1b: $F_{9,5,79} = 9.66, P < 0.001$) but not on incompatible trials (both $P$ values ≥ 0.25). This finding was qualified by an interaction between compatibility and flanker variance (Exp. 1a: $F_{1,7,64,91} = 9.72, P < 0.001$, Fig. S6D; Exp. 1b: $F_{1,9,66,79} = 9.67, P < 0.001$, Fig. 5E). 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 ($F_{1,7,61,9} = 10.61, P < 0.001$) and Exp. 1b ($F_{1,85,62,91} = 9.21, P < 0.001$). 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 conflict response alone. The fits for the gain model (colored circles) are shown superimposed upon the human data in Fig. 5D 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. S6A). 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 $\bar{X}$ fell equidistant to the category boundary, using instead a more complex design where they could vary independently around vertical at $(\pm 15°, \pm 30°, \pm 45°)$, and $X^i$ 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 $1/|X^i|$ (or equivalently, unscaled RTs) are indistinguishable when the target ($X^i$) and flanker mean $\bar{X}$ 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 $\bar{X}$, for example, when $X^i = 15$ and $\bar{X} = 30$ or $\bar{X} = 45$ (Fig. S4, 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 $\bar{X}$ 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 $= 3.76 \pm 3.9$ 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 $= 8.47 \pm 5.58$ 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
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.

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.

To assess the role of the dACC and interconnected regions in adaptive gain control, we conducted a new experiment in which $X^j$, $X^i$, and $X^c$ varied parametrically from trial to trial, rather than in a condition-wise 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...
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 begin 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.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 $|X'|$ in the dACC (peak: 2, 8, 54, $t_{19} = 9.06$, $P_{dpl} < 0.001$; $P_{dpl}$ denotes significance after correction for multiple comparisons using false discovery rate; Methods), as well as the anterior insula (AIC; peak: 34, 24, 2, $t_{19} = 10.19$, $P_{dpl} < 0.001$) and superior parietal lobe (SPL; peak: 22, −56, 46, $t_{19} = 8.19$, $P_{dpl} < 0.001$; see Fig. 7A). 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 $X'$ and $X$ (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. 7B). 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. 7C). Specifically, although BOLD responses were elevated when the $X'$ 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 $X$ 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: $t_{19} = 5.03$, $P = 0.0069$; AIC: $t_{19} = 2.82$, $P = 0.011$; SPL: $t_{19} = 2.28$, $P = 0.034$). No such modulation was observed when $X'$ 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 $|X'|$ 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. 8A; 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...
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 $1/X_j$ and RTs as competitive predictors in the model (GLM3), we still recovered a significant activation in the dACC (peak: 6, 20, 50; $t_{19} = 4.47; P_{FWE} = 0.046$) and AIC (left peak: $-30, 16, 10; t_{19} = 5.96; P_{FWE} = 0.038$; right peak: 34, 24, 6; $t_{19} = 5.05; P_{FWE} = 0.04$; 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 $x^j = 0$), the model predicts a larger output $|\hat{X}|$ 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 $|\hat{X}|$ but positively with $|\hat{X}|$. (GLM4; Fig. 8B; 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 in the model (dACC, AIC, and SPL) (19) then distract-
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 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 distractors 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 paralleling our findings here for trials with small \( X \) and large \( \tilde{X} \), 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 distractor’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 intervening 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. 2b: 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 \( (X') \), mean of flankers \( (\tilde{X}) \), and variability of flankers \( (\sigma^2) \). We can further designate trials as "congruent" when \( X' \) has the same sign as \( \tilde{X} \) or "incongruent" when \( X' \) has the opposite sign as \( \tilde{X} \). In Exp. 1, we thus use a 3 × 2 (flanker variability × congruency) within-participant factorial design. In Exp. 2, we introduced three levels of \( (\tilde{X}) \) (equal to six levels of signed \( X' \) for each category), three levels of \( (X') \) and three levels of \( (\sigma^2) \), resulting in 3 × 3 × 3 × 2 (\( X' \) × \( \tilde{X} \) × \( X' \) × \( \sigma^2 \)) within-participant factorial design with 54 conditions. A full list of \( X' \), \( \tilde{X} \), and \( X' \) 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 (\( \eta^2 \)) 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 10 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:

\[
RT_{\text{gain}} = \beta_0 + \beta_1 \cdot \frac{1}{|X'|}. 
\]  

This calculation of RT is equivalent to modeling the data with ballistic (noiseless) diffusion process, with the two additional parameters \( \beta_0 \) and \( \beta_1 \) encoding the nondetection time and the drift rate, respectively (fixed across conditions). We used a ballistic 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 \( \sigma^2 \) and \( \epsilon \) 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 \( A^{CCW} \):

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

\[
A^{CW} = g(X') \cdot (1 - w) + g(X') \cdot w \]  

\[
A^{CCW} = h(X') \cdot (1 - w) + h(X') \cdot w, \]  

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 \( X' \) (conflict model 2):

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

\[
A^{CW} = \frac{\sum g(X') \cdot (1 - w)}{\sum g(X') + h(X') \cdot w}, \]  

\[
A^{CCW} = \frac{\sum h(X') \cdot (1 - w)}{\sum h(X') + h(X') \cdot w}, \]  

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:

$$RT_{\text{cartsile}} = \beta^2 + \gamma^3 - C.$$  

[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 ($\sigma_{\text{max}}$) and a constant parameter ($\lambda$). Simulations of the model aimed at qualitatively recreating effects from the past literature (e.g., for Fig. 1) assumed a fixed $\sigma_{\text{max}}$ ($\sigma_{\text{max}} = 10$) and a fixed $\lambda$ ($\lambda = 5$) unless noted otherwise. We have also imposed a floor value for $\lambda$, 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 ($X^i$) as a function of flanker mean decision value $\vec{X} \in [-45, -44, \ldots, +45]$ and flanker SD $\sigma_X \in (3.7, 11, 15)$. For each variant of flanker mean decision values and flanker SD, $\vec{X}$ is computed using Eqs. 1 and 2. We then plot $\vec{X}$ against levels of flanker mean decision value and flanker SD in Fig. 1A.

Conflict effects. We computed a proxy of RT ($1/\vec{X}$) 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 there is a response association different from the flankers (Fig. 1B). In the simulation, flanker SD $\sigma_X^i$ is set to be 0 in both CO and RI conditions (i.e., we assumed flanker SD as $\sigma = 0$). In CO, $\vec{X}$ is equal to $\vec{X}$ (both are $+45$). In RI, $\vec{X}$ has the opposite sign to $\vec{X}$. Finally, we simulated the SI condition by assuming $\sigma_X^i$ is higher than 0 (i.e., $\sigma_X^i = 5$); individual flankers are variable but $\vec{X}$ remained the same as $\vec{X}$. We assumed a higher maximum tuning width ($\sigma_{\text{max}} = 15$) 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, $Z \in [-45, -44, \ldots, +20]$. $\tau$ was assumed to be 10 in this simulation. In SI Appendix, Fig. 52, we computed the model output associated with the highest-valued choice-relevant alternative ($X$) and the next-best alternative ($Y$), 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 ($X - Y$), $Y$, a quantity proportional to choice probability in Fig. 1C.

Decoy effects (multiatribute). We simulated the influence of a distractor ($Z$) 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 indifference (i.e., parallel to the indifference line). For illustration, we use $X = [15, 10]$ and $Y = [10, 15]$. We implemented the adaptive gain parameter coding model as follows:

$$a_i = a_{\text{max}} - f(q_i, 2\sigma_i, \epsilon) \cdot \epsilon.$$  

[7]

Like in the single-attribue case, we assume that the inverse Gaussian tuning distribution is centered at the decoy attribute value, $Z_d$, with an SD $\epsilon = 30$, and maximum tuning width $\sigma_{\text{max}} = 10$ and an extra free parameter: tuning width range $\tau = 440$. Each neuron responds to the value of the attributes $X_0$ and $Y_0$, resulting two hills of activity across the population of simulated neurons. The activity is decoded to subjective estimates difference for each of the two option values on attribute $P(X_0)$ and $Y_0$ 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:

$$X = X_0 + X_0' \times Y = Y_0 + Y_0' \times X.$$  

Fig. 2 shows the relative subjective value difference ($X - Y$) 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 $\sigma_{\text{max}}$, ranging from 8 to 16 with 40 linearly spaced values). Subjects with a high $\sigma_{\text{max}}$ would have overall more broadly tuned tuning curves than subjects with a low $\sigma_{\text{max}}$ (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. 3E (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. 3A). 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 ($X - Y$) is corrupted by a noise term $\epsilon$, that is normally 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 $X$ and $Y$. 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 anti-correlation 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: $X = (15, 10)$, $Y = (10, 15)$, $Z = ([13,5]$; $\sigma_{\text{max}} = 20); \epsilon = 10$, and $\tau = 400$. After obtaining decoded subjective estimates $X, Y, Z$, we then add noise to these subjective estimates by sampling from zero-mean Gaussian distribution with differing SDs ($\sigma_{\text{max}}$, 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 $\epsilon = (20.5)$. For this analysis, we used a slightly different set of parameters: $\sigma_{\text{max}} = 10$, $\epsilon = 140$, and $\epsilon = 0.08-0.3$. These parameters allow us to recreate mostly faithfully the compromise choice effect (whether under time pressure shown in the Pettit et al. 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. 53.

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 $\epsilon = 10$. On each trial, decision values for two stimuli ($X^i$ and $X^j$) 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:

$$D = X_{\text{max}} - X_{\text{min}}$$

[8]

$$CP = \frac{1}{1 + e^{k}}$$

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 orthogonization 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 < 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 disentangled each parametric modulator (i) target decision values $X^i$ and (ii) flanker mean feature value $\overline{X}$ into four quartiles bins and included a total of 4 × 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 $1/(\sqrt{X})$ (a full list of active voxels for $1/(\sqrt{X})$ is 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
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 009974/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 University 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. Logan GD (1996) The CODE theory of visual attention: An integration of space-based and object-based attention. Neuron 17:463–470.
6. Tsetsos K, Usher M, Chater N (2010) Preference reversal in multiattribute choice. Psychol Rev 117:1275–1293.
7. Tversky A (1972) Elimination by aspects: A theory of choice. Psychol Rev 79:281–299.
8. Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. Psychol Rev 108:624–652.
9. Bundesen C (1998) A computational theory of visual attention. Philos Trans R Soc Lond B Biol Sci 353:1271–1281.
10. Heeger DJ (1992) Normalization of cell responses in cat striate cortex. Vision Res 32:179–198.
11. Tsetsos K, Wu T, Dufford AJ, Mackie MA, Egan LJ, Fan J (2016) The capacity of cognitive control protected from a perceptual decision-making task. Sci Rep 6:34025.
12. Padoa-Schioppa C (2009) Range-adapting representation of economic value in the human visual system. Nature 228:37–39.
13. Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representations. Neuron 29:14004–14014.
14. Dayan P, Summerfield C (2014) Priming by the variability of visual context effects. Nature 514:224–225.
15. Schwartz O, Sejnowski TJ, Dayan P (2009) Perceptual organization in the tilt illusion. Proc Natl Acad Sci USA 106:17199–17204.
16. Heeger DJ (1993) Normalization of cell responses in cat striate cortex. Vision Res 33:2242–2253.
17. Soltani A, De Martino B, Camerer C (2012) A range-normalization model of context effects. J Cogn Neurosci 24:1193–1216.
18. Schwartz O, Sejnowski TJ, Dayan P (2009) Perceptual organization in the tilt illusion. Proc Natl Acad Sci USA 106:17199–17204.
19. Dayan P, Summerfield C (2014) Priming by the variability of visual context effects. Nature 514:224–225.
20. Michael E, de Gardelle V, Summerfield C (2014) Priming by the variability of visual context effects. Nature 514:224–225.
21. Simonson I (1989) Choice based on reasons: The case of attraction and compromise decoys in choice. J Exp Psychol Gen 118:562–582.
22. Padoa-Schioppa C (2009) Range-adapting representation of economic value in the human visual system. Nature 228:37–39.
23. Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representations. Neuron 29:14004–14014.
24. Dayan P, Solomon JA (2010) Selective Bayes: Attentional load and crowding. Vision Res 50:2248–2250.
25. Michael E, de Gardelle V, Summerfield C (2014) Priming by the variability of visual context effects. Nature 514:224–225.
26. Kolling N, et al. (2016) Value, search, persistence and model updating in anterior cingulate cortex. J Neurosci 36:95–109.
27. Heeger DJ (1992) Normalization of cell responses in cat striate cortex. Vision Res 32:179–198.
28. Schwartz O, Sejnowski TJ, Dayan P (2009) Perceptual organization in the tilt illusion. Proc Natl Acad Sci USA 106:17199–17204.
29. Wu T, Dufford AJ, Mackie MA, Egan LJ, Fan J (2016) The capacity of cognitive control protected from a perceptual decision-making task. Sci Rep 6:34025.
30. de Gardelle V, Summerfield C (2011) Robust averaging during perceptual judgment. Proc Natl Acad Sci USA 108:13341–13346.
31. Michael E, de Gardelle V, Summerfield C (2017) Robust averaging protects decisions from noise in neural computations. PLoS Comput Biol 13:e1005722.
32. Solomon JA, Morgan MJ (2006) Stochastic re-calibration: Contextual effects on perceived tilt. Proc Biol Sci 273:2681–2686.
33. Simonsen I (1989) Choice based on reasons: The case of attraction and compromise decoys in choice. J Exp Psychol Gen 118:562–582.