Automatic integration of confidence in the brain valuation signal

Maël Lebreton[1,2], Raphaëlle Abitbol[1–3], Jean Daunizeau[1,2] & Mathias Pessiglione[1,2]

A key process in decision-making is estimating the value of possible outcomes. Growing evidence suggests that different types of values are automatically encoded in the ventromedial prefrontal cortex (VMPFC). Here we extend this idea by suggesting that any overt judgment is accompanied by a second-order valuation (a confidence estimate), which is also automatically incorporated in VMPFC activity. In accordance with the predictions of our normative model of rating tasks, two behavioral experiments showed that confidence levels were quadratically related to first-order judgments (age, value or probability ratings). The analysis of three functional magnetic resonance imaging data sets using similar rating tasks confirmed that the quadratic extension of first-order ratings (our proxy for confidence) was encoded in VMPFC activity, even if no confidence judgment was required of the participants. Such an automatic aggregation of value and confidence in a same brain region might provide insight into many distortions of judgment and choice.

Decision theory in its elementary form assumes that, when faced with a choice, individuals first assign subjective values to possible outcomes of available actions and then select the action that leads to the most valuable outcomes[1,2]. Estimating the value of potential world states is therefore a key step of the decision-making process. At a psychological level, subjective value can be understood as anticipated pleasure and can be measured using pleasantness or desirability rating. A growing body of evidence from both animals and humans suggests that subjective value, whether measured with rating or inferred from choice, is encoded in the VMPFC[3–5]. Two important properties of the brain's value signal have been uncovered: generality and automaticity. Generality means that the brain value signal can rank seemingly incommensurable items on a common scale. Indeed, the VMPFC has been found to reflect subjective values of faces, houses, cars, paintings, music, food, money, social donation and so forth[6–10]. Automaticity means that subjective values are encoded in VMPFC activity even when unnecessary for the ongoing task—for example, during passive viewing of visual stimuli or during the rating of dimensions independent from value. These automatic value signals were predictive of the choices probed after neuroimaging measurements[9,11–13].

Yet decoding values from VMPFC activity might not be straightforward, since this region has been found to incorporate other factors. In particular, several studies in both rats and humans have reported that orbitofrontal or ventromedial prefrontal regions encode confidence in having made the best choice[14–16]. Borrowing from established theoretical work in the psychophysics of perceptual decisions[17,18], these studies have modeled confidence as the temporal integration of the (noisy) difference in value between choice options, through either linear accumulation or nonlinear competition between attractors. Here we generalize the notion of confidence-encoding in the VMPFC to all situations that involve a judgment on any dimension (value and others), and not only to those implying a choice between alternative options. In brief, we think of confidence as a second-order judgment on the correctness of a first-order judgment that can be either a choice or a rating[19,20]. In this perspective, the hypothesis that confidence is encoded in the VMPFC remains compatible with the function attributed to this region: namely, computing subjective value. This is simply because being accurate is valuable: it is the implicit goal of any judgment task. Considering confidence as a value judgment also suggests that the neural confidence signal might share the properties of the neural value signal: it should be general (that is, elicited by any kind of first-order judgment) and automatic (that is, elicited even when no confidence judgment is explicitly required).

We tested these hypotheses using a combination of published and original data sets acquired in healthy human participants. We reasoned that confidence should vary as a non-monotonic (U-shaped) function of first-order judgments. This is essentially because participants tend to stay on the middle of the rating scale when they have no idea how to respond and move to the extremes when they have more information. Here we provide empirical evidence, using two behavioral studies, for the U-shaped (or quadratic) relationship between second-order confidence rating and four sorts of first-order ratings: pleasantness, age, desirability and probability. We also show, in three functional magnetic resonance imaging (fMRI) studies, that the quadratic extension of all first-order judgments is encoded in VMPFC activity, even in the absence of explicit confidence ratings. We conclude that response confidence is automatically aggregated with the usual stimulus value signal in the VMPFC.

RESULTS

Computational analysis of rating tasks

To provide a normative account of our working hypothesis, we suggest a computational decomposition of rating tasks (see Online Methods). Starting from a decision-theoretic perspective, we assume that

[1]Motivation, Brain and Behavior team, Centre de NeuroImagerie de Recherche (CENIR), Institut du Cerveau et de la Moelle épinière (ICM), Paris, France. [2]INSERM UMRS 975, CNRS UMR 7225, Université Pierre et Marie Curie UPMC-Paris 6 UMR 1127, Paris, France. [3]Centre d'Economie de la Sorbonne, Université Paris 1-Panthéon-Sorbonne, Paris, France. Correspondence should be addressed to M.P. (mathias.pessiglione@gmail.com).

Received 24 April; accepted 19 June; published online 20 July 2015; doi:10.1038/nn.4064

participants aim at minimizing the mismatch between their overt rating and their internal judgment. Any potential rating thus induces a subjective feeling about being right or wrong—that is, an estimation of accuracy that guides the selection of the overt response in the task. This motivates our formal definition of confidence: namely, the expected judgment accuracy. Estimating the expected accuracy is not trivial, for the following two reasons. First, subjects may be uncertain about their judgment. Thus the information that subjects have about their internal judgment may be captured not by a number but by a probability distribution. Second, their internal judgment may not be naturally expressed in the metric imposed by the rating scale. That is, it may have to be mapped onto this external bounded scale through a monotonic function that preserves preference ordering. Taken together, these difficulties imply that subjects must deal with degraded information about how to respond, expressed in terms of a probability distribution on mapped judgments (**Fig. 1a**). Analytical decomposition of the model shows that rating and confidence should correspond to the first and second-order moments of this distribution, respectively. This allows deriving rating and confidence from the mean and variance of the internal probability distribution (**Fig. 1b**). Critically, the model predicts that, irrespective of the actual internal judgments, these two quantities are not independent: confidence is a quadratic function of rating (**Fig. 1c**). In summary, our computational analysis explains, under the intuitive assumption that subjects intend to express their judgment as accurately as possible, both why confidence possesses an intrinsic value and why it is minimal for midscale ratings.

Study 1a: relationship between second-order (confidence) and first-order (age, pleasantness) ratings

To establish that a confidence rating is automatically generated and aggregated with the brain value signal, it was first necessary to demonstrate a systematic statistical relationship between rating and confidence. We added a confidence rating task on top of a pleasantness rating task, for which we already had an fMRI data set[9], and tested a new group of healthy participants (n = 18). A visual stimulus from three possible categories (faces, houses or paintings) was presented on every trial and participants were asked to estimate first the stimulus pleasantness on a discrete scale (from −10 to +10) and then their confidence in their pleasantness rating on a continuous analog scale (from not at all to totally confident) (**Fig. 2a**).

Consistent with our predictions, behavioral data showed positive U-shaped association between confidence and pleasantness ratings (**Fig. 2b**, top). We used robust multiple regression across trials at the individual level, in order to explain confidence ratings with both linear and squared pleasantness ratings. A random-effect analysis at the group level (one-sample t tests on individual robust regression estimates (β)) confirmed that confidence varied not with pleasantness but with squared pleasantness (linear: $\beta = -0.3 \pm 0.04$, $t_{17} = -0.80$, $P = 0.42$; quadratic: $\beta = 0.35 \pm 0.04$, $t_{17} = 9.94$, $P < 10^{-6}$). The same pattern (with opposite sign) was observed in the reaction time measured for pleasantness rating, as the time between stimulus onset and first key press (linear: $\beta = 0.01 \pm 0.03$, $t_{17} = 0.47$, $P = 0.65$; quadratic: $\beta = -0.12 \pm 0.02$, $t_{17} = -5.42$, $P = 4.6 \times 10^{-5}$). Consequently, confidence and reaction time were linearly correlated (random-effect analysis on individual Pearson correlation $\rho = -0.12 \pm 0.04$, $t_{17} = -3.44$, $P = 0.0031$). The model predicts that when internal judgments are more uncertain the quadratic effect should be deeper, and not simply shifted toward lower confidence levels (**Fig. 1c**). To test this prediction, we divided each rating bin into tertiles (low, medium and high confidence). The weight of the quadratic term was indeed significantly greater in the low- than in the high-confidence tertiles (high confidence: $\beta = 0.20 \pm 0.02$; low confidence: $\beta = 0.67 \pm 0.06$; difference: $t_{17} = 9.21$, $P < 10^{-6}$).

To examine whether this statistical relationship with subjective value could be extended to an orthogonal first-order judgment on a more objective dimension, we added the same confidence rating task on top of an age rating task that we also had used in our previous fMRI study[9] and tested another group of healthy participants (n = 22). A visual stimulus from the same three possible categories (faces, houses or paintings) was presented on every trial and participants were asked to estimate first the stimulus age on a discrete scale (from 20 to 50 years old for faces, date of creation between 1400 and 2000 for painting, and date of construction between 1700 and 2000 for houses) and then their confidence in their age rating on a continuous analog scale (from not at all to totally confident) (**Fig. 2a**).

As observed with pleasantness ratings, we found a positive U-shaped association between confidence and age ratings (**Fig. 2b**, bottom).

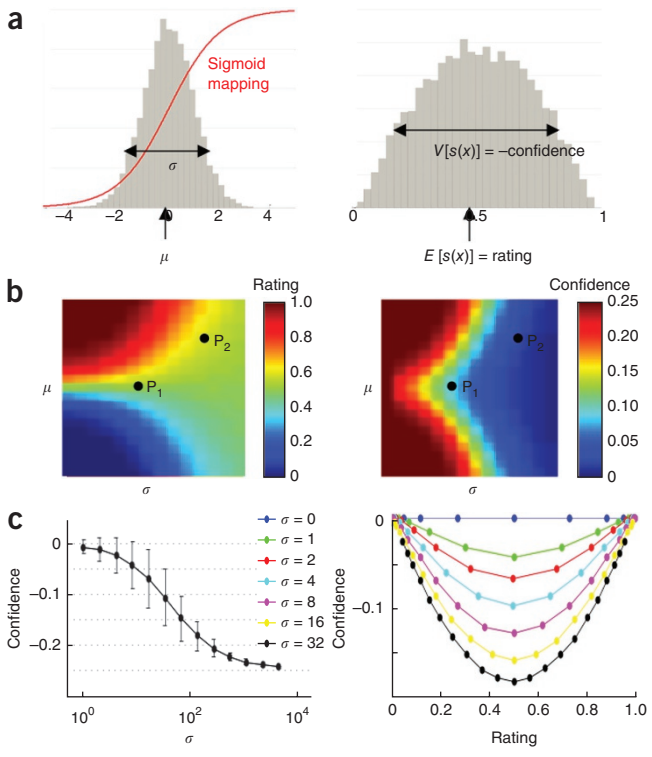

Figure 1 Model simulations. (**a**) Left, the agent's probability distribution $p(x)$ over her internal judgment x, with mean μ and variance σ (which captures judgment uncertainty). The sigmoidal projection mapping is shown in red. Right, the induced probability distribution $p[s(x)]$ over projected judgments $s(x)$, with first-order moment $E[s(x)]$ and second-order moment $V[s(x)]$). (**b**) Left, predicted rating $\hat{r} = E[s(x)]$ is shown with color code as a function of mean μ (y axis) and uncertainty σ (x axis) of internal judgments. Points lying on iso-rating curves show that midrange overt ratings mix both neutral judgments (P_1) and uncertain judgments (P_2). Right, predicted confidence $\hat{q} = -V[s(x)]$ is shown with color code as function of μ (y axis) and σ (x axis). (**c**) Left, marginal relationship between predicted confidence $\hat{q}$ (y axis) and judgment uncertainty σ, after averaging over judgment mean μ. The variability induced by μ is shown by error bars, which depict 1 s.d. around the mean. Confidence is monotonically related to judgment uncertainty. Right, predicted confidence $\hat{q}$ (y axis) is shown as a function of predicted rating $\hat{r}$ (x axis), for different levels of judgment uncertainty (σ ranges from 0 to 32). The quadratic relationship between confidence and rating is more pronounced at greater σ.

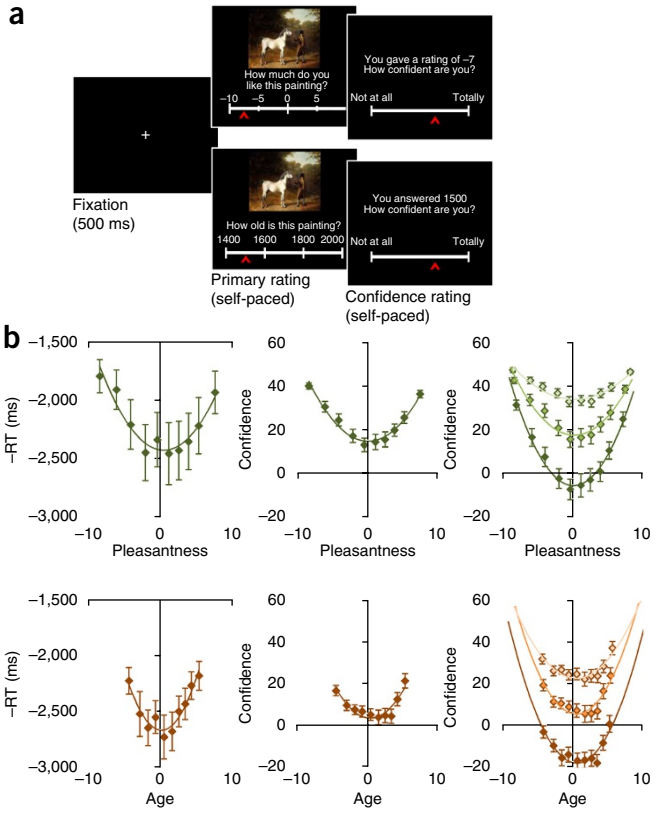

Figure 2 Relationship between confidence and stimulus pleasantness or age rating (study 1a). (**a**) Task design. Successive screens displayed in one trial are shown from left to right, with their durations. Subjects had first to estimate the pleasantness or the age of a visual stimulus (face, house or painting) on a −10 to 10 scale. Then they were asked to estimate on a continuous scale how confident they were in their pleasantness or age rating, which was given on the screen. (**b**) Behavioral results. Both reaction time (left) and confidence rating (middle) vary as U-shaped functions of pleasantness (top) or age (bottom) rating. Reaction time is the interval between stimulus onset and first key press. Trials were grouped in ten bins of ascending ratings sorted at the individual level and then averaged across individuals. At right, each rating bin was divided into tertiles of confidence (light to dark color). Error bars indicate intersubject s.e.m. Solid lines indicate the best second-order polynomial fit.

In other words, confidence ratings varied not linearly but quadratically with age ratings (linear: $\beta = 0.06 \pm 0.04$, $t_{21} = 1.55$, $P = 0.14$; quadratic: $\beta = 0.22 \pm 0.03$, $t_{21} = 6.63$, $P = 1.5 \times 10^{-6}$). The same pattern (with opposite sign) was observed in the reaction time measured for age rating (linear: $\beta = -0.02 \pm 0.02$, $t_{21} = -0.90$, $P = 0.37$; quadratic: $\beta = -0.08 \pm 0.01$, $t_{21} = -4.71$, $P = 1.2 \times 10^{-4}$). Consequently, confidence and reaction time were linearly correlated ($\rho = -0.08 \pm 0.02$, $t_{21} = -3.58$, $P = 0.0018$). We also observed a trend toward a deepening of quadratic functions with confidence tertiles (high confidence: $\beta = 0.17 \pm 0.02$; low confidence: $\beta = 0.25 \pm 0.06$; difference: $t_{21} = 1.93$, $P = 0.067$), although it was less clear than with pleasantness ratings, probably owing to sampling issues (confidence was globally lower in age relative to pleasantness ratings).

The correlation between reaction time and confidence level corresponds to the intuition that harder (more uncertain) judgments take longer, whatever the domain. It suggests that the uncertainty expressed in confidence ratings is already available during first-order judgment, before elicitation of the second-order judgment. Thus, these behavioral results made it possible to search for neural confidence signals during both pleasantness and age rating tasks, without having to ask subjects to rate their confidence. However, we noted that the quadratic link was stronger with confidence than with reaction time. To compare their capacity to explain confidence level, we included reaction time together with response time (delay between stimulus onset and last key press) and linear and quadratic functions of ratings in a same regression model. The only significant predictor across rating tasks was the quadratic term (pleasantness: $\beta = 0.33 \pm 0.03$, $t_{17} = 10.15$, $P < 10^{-6}$; age: $\beta = 0.23 \pm 0.03$, $t_{21} = 7.58$, $P < 10^{-6}$). This discards the possibility that confidence could be a simple readout of reaction time, and suggests instead that confidence and reaction time both arise from the same underlying uncertainty.

Study 1b: neural representation of confidence in pleasantness and age rating

We reanalyzed fMRI data from a previous study[9] that demonstrated linear encoding of pleasantness rating in a set of brain regions including the VMPFC. This study employed the same stimuli (faces, houses and paintings) and the same pleasantness and age rating tasks as in study 1a, except that no confidence rating was ever asked of participants (**Fig. 3a**).

All age-rating and pleasantness-rating sessions were modeled with a boxcar function covering painting display modulated by a second-order polynomial expansion of pleasantness or age rating (with linear and quadratic terms in separate regressors). Model estimates first confirmed that pleasantness ratings were linearly encoded in the VMPFC (**Fig. 3b**), but age ratings were not, even at a permissive threshold (uncorrected $P_{unc} < 0.01$). At the whole-brain level, linear activation with pleasantness was restricted to a VMPFC cluster specifically (peak MNI coordinates [−2 52 −2], $k = 214$ voxels) when using a statistical threshold family-wise error corrected for multiple comparisons at the voxel level ($P_{FWE\_vox} < 0.05$).

To test whether pleasantness and confidence were encoded in the same brain regions, we performed a conjunction analysis between the linear and quadratic expansions of pleasantness ratings, as well as the quadratic expansion of age rating. This conjunction isolated several functional clusters (at a threshold of $P_{FWE\_clu} < 0.05$) in the visual and frontal cortex, including the VMPFC (**Fig. 3c**). To confirm that confidence was indeed encoded in the VMPFC valuation region, we extracted regression estimates (β values) obtained for the linear and quadratic terms from the VMPFC region of interest (ROI), defined by linear activation with pleasantness rating at the group level (**Fig. 3b**). Linear and quadratic estimates denoted a similar effect size and were both highly significant (one-sample t test on individual β values, linear: $\beta = 0.20 \pm 0.02$, $t_{19} = 10.75$, $P < 10^{-6}$; quadratic: $\beta = 0.16 \pm 0.02$, $t_{19} = 7.98$, $P < 10^{-6}$). Note that the ROI was selected for encoding pleasantness rating, so testing the linear relation with pleasantness is just confirmatory, whereas the quadratic relation is entirely independent of the selection criterion. We also tested in the same VMPFC region the linear and quadratic regression with age. Only the quadratic term was significantly encoded (linear: $\beta = -0.03 \pm 0.03$, $t_{19} = 1.04$, $P = 0.31$; quadratic: $\beta = 0.07 \pm 0.03$, $t_{19} = 2.30$, $P = 0.033$). These results were similar whether or not response time was included in the regression model, even if not orthogonalized with respect to the quadratic regressor.

These results are illustrated by plotting the VMPFC signal obtained for nine bins of ascending pleasantness or age (**Fig. 3d**). While the curve obtained for pleasantness appears as an addition of linear and quadratic terms, the curve obtained for age is purely quadratic (U-shaped). Thus, fMRI data showed that confidence, modeled as

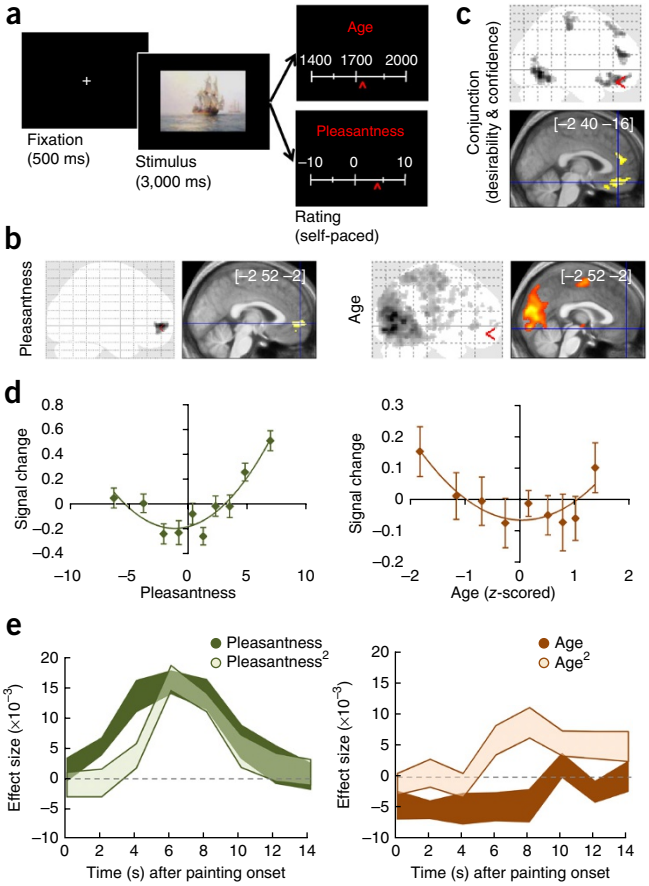

Figure 3 Neural integration of confidence in stimulus pleasantness and age rating (study 1b). (**a**) Task design. Successive screens displayed in one trial are shown from left to right, with their durations. Subjects first viewed a stimulus (face, house or painting) and then had to rate either its pleasantness or its age. (**b**) Statistical parametric maps of activation with pleasantness (cluster generating threshold $P_{FWE\_vox} < 0.05$ and $k > 100$ voxels) and age ratings (uncorrected voxel threshold $P_{unc\_vox} < 0.01$). The color code on glass brains (gray to black) and sagittal slices (red to yellow) indicates the *t*-value of clusters that surpassed the statistical threshold. The [*x y z*] coordinates of the maxima (blue crosshairs and corresponding red pointers) refer to Montreal Neurological Institute (MNI) space. (**c**) Conjunction among linear correlation with pleasantness, quadratic correlation with pleasantness and quadratic correlation with age (cluster generating threshold, $P_{unc\_vox} < 0.001$; cluster family-wise correction, $P_{FWE\_clu} < 0.05$). (**d**) Canonical ROI analysis. Trials were grouped for each participant in ten bins of ascending pleasantness (left) or age (right), averaged at the population level and plotted against the BOLD signal estimated in the VMPFC ROI (isolated from group-level linear correlation with pleasantness; **b**, left). Age was *z*-scored because the age rating scale varied with stimulus category (face, house, painting). Error bars indicate intersubject s.e.m.; solid lines indicate the best second-order polynomial fit (including both linear and quadratic terms). (**e**) Time-resolved ROI analysis. A FIR model estimated the effect size of the linear and quadratic expansions of pleasantness (left) and age (right) ratings on the VMPFC signal, every 2 s after painting onset (time 0). The VMPFC ROI is the same as above. Shaded areas represent confidence intervals (means ± intersubject s.e.m.).

squared first-order ratings (pleasantness or age), is encoded in a VMPFC region that also encodes value (linear variation with pleasantness but not age). Integration of confidence in this valuation region can be considered automatic, since no confidence judgment was ever asked of participants in the fMRI study. To further illustrate this finding, we estimated a parametric finite-impulse response (FIR) model, in which the blood oxygen level–dependent (BOLD) signal was fitted every 2 s with both the linear and quadratic extensions of pleasantness or age ratings. This procedure allowed us to estimate the time course of the effect size, which was similar for the linear and quadratic predictors (**Fig. 3e**).

Study 2: neural representation of confidence in desirability rating

The same pattern of results, in the same VMPFC ROI, was also observed in another independent fMRI data set[21] previously recorded in a group of 19 participants during a desirability rating task (**Fig. 4**). The task involved subjects rating the desirability of objects (toys, tools or food items) featured in a video where an actor could or could not perform an action directed to this object (**Fig. 4a**). Using parametric modulation of activity modeled as a boxcar over video viewing, our previous analysis had shown that, irrespective of actions, desirability ratings were reflected in the VMPFC, as well as in a large occipital cluster (**Fig. 4b**, $P_{FWE\_clu} < 0.05$). We have now added the quadratic extension of desirability as a second parametric modulator in our general linear model, and found again that the VMPFC encoded a conjunction of first-order (desirability) and second-order (confidence) value judgments (**Fig. 4c**, $P_{FWE\_clu} < 0.05$). Regression estimates were extracted from our VMPFC ROI (independently defined from the previous study) and tested at the group level to confirm that they had a similar effect size and were both highly significant (one-sample *t* tests on individual regression estimates; linear: $\beta = 0.48 \pm 0.13$, $t_{17} = 3.46$, $P = 0.0033$; quadratic: $\beta = 0.45 \pm 0.13$, $t_{17} = 3.61$, $P = 0.0023$). The aggregation of linear and quadratic terms was also observable when plotting the signal change as a function of ascending bins (**Fig. 4d**) or the time course of effect sizes (**Fig. 4e**).

This external replication suggests that integration of confidence in regions isolated for encoding value is a robust phenomenon. It can be extended to judgments of desirability—that is, an anticipated value—as opposed to pleasantness, which might be considered an experienced value.

Study 3a: relationship between second-order (confidence) and first-order (desirability, probability) ratings

An obvious confounding factor for confidence is salience, which is generally conceived as unsigned stimulus value. The idea is that very pleasant and very unpleasant stimuli are both salient compared to neutral stimuli. Thus, salience would also vary as a U-shaped (or V-shaped) function of stimulus value. The fact that the VMPFC also signaled confidence in age rating could be taken as evidence against a confound with salience, but one could argue that very old and very new stimuli are more salient than middle-aged ones. We also noted that the results obtained with age rating were globally less clear-cut than with pleasantness rating. We therefore explored another subjective judgment, for which salience and confidence would not covary. We chose probability (likelihood), because salience is not expected to vary as a U-shaped function of probability: by definition very improbable events are salient but very probable events are not. In the new task designed for study 3, participants had to rate either the desirability or the probability of future events (prospects; see examples in **Supplementary Table 1**). Participants were asked to rate either the desirability or the probability of factual prospects from various domains (culture, sport, society, economics, diplomacy, science, technology and so forth): for example, France wins the next World Football cup. A first group of healthy participants ($n = 18$) performed a version of this task (**Fig. 5a**) that included rating confidence in first-order judgments (desirability or probability rating). Our hypothesis was that the VMPFC would not encode probability, because probability

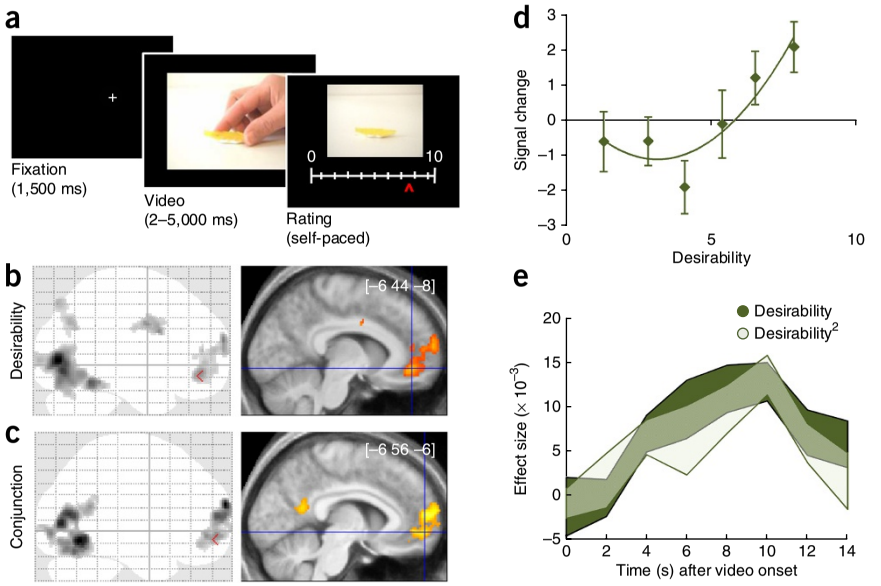

Figure 4 Neural integration of confidence in stimulus desirability rating (study 2). (**a**) Task design. Successive screens displayed in one trial are shown from left to right, with their durations. Subjects first watched a video depicting a stimulus (tool, toy or food item), with or without and actor pointing to this stimulus, and then had to judge its desirability on a scale of 0 to 10. (**b**) Statistical parametric map of activations with desirability ratings (cluster generating threshold, $P_{unc\_vox} < 0.001$; cluster family-wise correction, $P_{FWE\_clu} < 0.05$). The color code on glass brains (left maps) and axial slices (right maps) indicates the statistical significance of clusters that surpassed the thresholds. The [*x y z*] coordinates of the maxima indicated by the blue crosshairs and corresponding red pointers refer to Montreal Neurological Institute (MNI) space. (**c**) Conjunction between the linear and the quadratic expansions of the parametric desirability ratings (cluster generating threshold, $P_{unc\_vox} < 0.005$; cluster family-wise correction, $P_{FWE\_clu} < 0.05$). (**d**) Canonical ROI analysis. Trials were grouped per participants in six bins of ascending desirability, averaged at the population level and plotted against the BOLD signal estimated in the VMPFC ROI (isolated from group-level linear correlation with pleasantness; **Fig. 3b**, left). Error bars indicate intersubject s.e.m.; solid lines indicate the best second-order polynomial fit (including both linear and quadratic terms). (**e**) Time-resolved ROI analysis. A FIR model estimated the effect size of the linear and quadratic expansions of desirability ratings on the VMPFC signal every 2 s after stimulus onset (time 0). The VMPFC ROI is the same as above. Shaded areas represent confidence intervals (means ± intersubject s.e.m.).

is, at least theoretically, orthogonal to value. As seen for pleasantness and age, the VMPFC would encode confidence in both desirability and probability. This would confirm our interpretation since it would show that the VMPFC signals the probability of being correct (that is, confidence) but not the probability of external events.

Replicating the result of study 1a regarding subjective value (**Fig. 2b**), we found that confidence ratings were well accounted for by squared desirability ratings (**Fig. 5b**), but not by the linear regressor (one-sample *t* tests on individual regression estimates; linear: $\beta = -0.01 \pm 0.04$, $t_{17} = -0.25$, $P = 0.80$; quadratic: 0.45 ± 0.05, $t_{17} = 9.27$, $P < 10^{-6}$). Extending this result, we found that confidence ratings were also explained by squared probability ratings, but only marginally by the linear regressor (quadratic: $\beta = 0.47 \pm 0.04$, $t_{17} = 10.94$, $P < 10^{-6}$; linear: -0.15 ± 0.07, $t_{17} = -2.16$, $P = 0.045$). The unexpected linear trend was likely due to the fact that probability ratings were not exactly centered on 50% but biased toward lower estimates (paired *t*-test against 50%: rating = 0.42 ± 0.1, $t_{17} = -4.45$, $P = 3.5 \times 10^{-4}$). In line with model predictions (**Fig. 1c**) and the results of study 1a (**Fig. 2b**), the quadratic functions relating first-order rating to confidence level were deeper for low-confidence than for high-confidence tertiles, for both desirability (high confidence: $\beta = 0.30 \pm 0.0$; low confidence: $\beta = 0.81 \pm 0.07$; difference: $t_{17} = 7.40$, $P = 3.5 \times 10^{-6}$) and probability (high confidence: $\beta = 0.34 \pm 0.04$; low confidence: $\beta = 0.70 \pm 0.08$; difference: $t_{17} = 4.45$, $P = 3.5 \times 10^{-4}$).

Also in accordance with the results of study 1a, reaction time for first-order judgments was negatively correlated with the confidence assigned to these judgments (desirability: $\rho = -0.20 \pm 0.03$, $t_{17} = -7.52$, $P < 10^{-6}$;

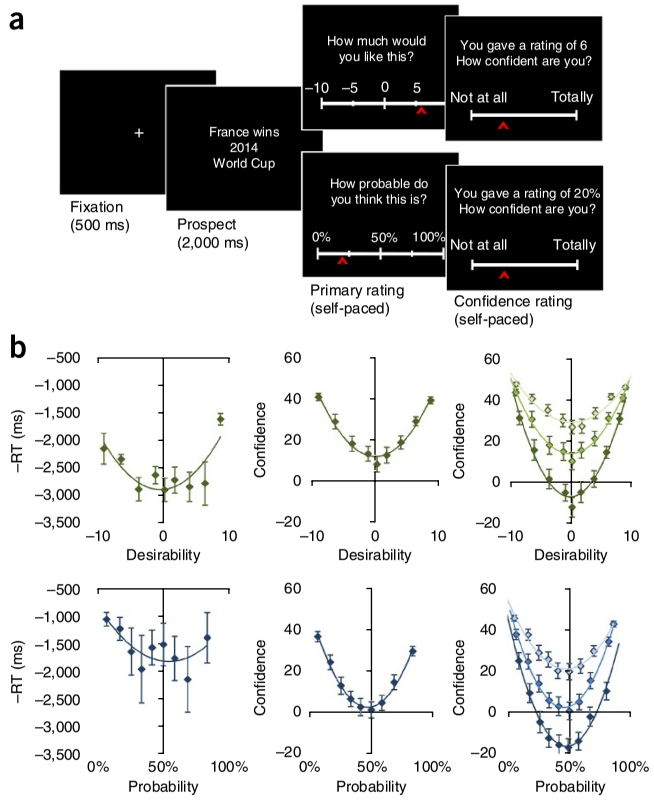

Figure 5 Relationship between confidence and prospect desirability or probability (study 3a). (**a**) Task design. Successive screens displayed in one trial are shown from left to right, with their durations. Subjects first read a prospect and then had to estimate either its desirability (between –10 and 10) or its probability of occurrence in the near future (between 0 and 100%). Then, participants were asked to estimate on a continuous scale how confident they were in their first-order rating, which was given on the screen. (**b**) Behavioral results. Both reaction time (left) and confidence rating (middle) vary as a U-shaped function of either desirability (top) or probability (bottom) rating. Reaction time is the interval between stimulus onset and first key press. Trials were grouped per participants in ten bins of ascending desirability or probability sorted at the individual level and then averaged across individuals. At right, each rating bin was divided into tertiles of confidence (light to dark color). Error bars indicate intersubject s.e.m. Solid lines indicate the best second-order polynomial fit.

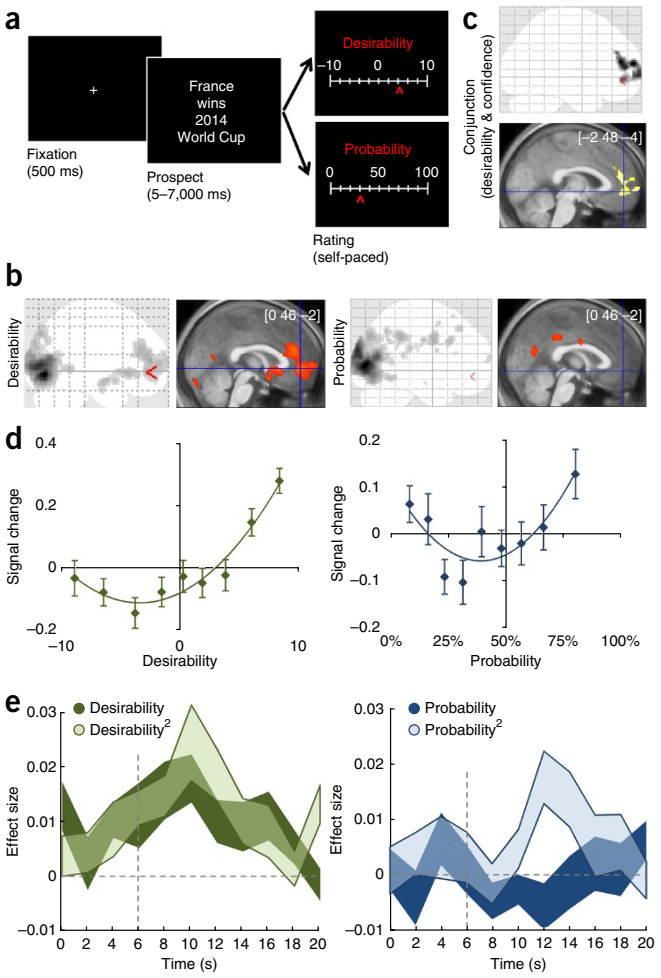

Figure 6 Neural integration of confidence in prospect desirability and probability rating (study 3b). (**a**) Task design. Successive screens displayed in one trial are shown from left to right, with their durations. Subjects first read a prospect and then had to rate either its desirability (between −10 and 10) or its probability of occurrence in the near future (between 0 and 100%). (**b**) Statistical parametric maps of activation with desirability (cluster generating threshold, $P_{unc\_vox} < 0.001$; cluster family-wise correction, $P_{FWE\_clu} < 0.05$) and probability (uncorrected voxel threshold, $P_{unc\_vox} < 0.01$) ratings. The color code on glass brains (gray to black) and sagittal slices (red to yellow) indicates the *t*-value of clusters that surpassed statistical threshold. The [*x y z*] coordinates of the maxima indicated by the blue crosshairs and corresponding red pointers refer to Montreal Neurological Institute (MNI) space. (**c**) Conjunction among linear correlation with desirability, quadratic correlation with desirability and quadratic correlation with probability (cluster generating threshold, $P_{unc\_vox} < 0.005$; cluster family-wise correction, $P_{FWE\_clu} < 0.05$). (**d**) Canonical ROI analysis. Trials were grouped for each participants in ten bins of ascending desirability (left) and probability (right), averaged at the population level and plotted against the BOLD signal estimated in the VMPFC ROI (isolated from group-level linear correlation with pleasantness; **Fig. 3b**, left). Error bars indicate intersubject s.e.m.; solid lines indicate the best second-order polynomial fit (including both linear and quadratic terms). (**e**) Time-resolved ROI analysis. A FIR model estimated the effect size of the linear and quadratic expansions of desirability (left) and probability (right) ratings on the VMPFC signal every 2 s after stimulus onset for linear expansions and every two 2 s after a starting point aligned to scale onset minus 6 s (corresponding to stimulus onset on average) for quadratic expansions. The VMPFC ROI is the same as above. Shaded areas represent confidence intervals (means ± intersubject s.e.m.).

probability: $\rho = -0.17 \pm 0.2$, $t_{17} = -8.66$, $P < 10^{-6}$). Again, this suggests that confidence rating was based on a process that was occurring during first-order judgment and that could therefore be captured with fMRI even if no confidence judgment was explicitly required. We also verified with this new data set that, when squared rating and reaction time were incorporated into the same regression model, only squared rating was a significant predictor of confidence level across rating tasks (desirability: $\beta = 0.44 \pm 0.05$, $t_{17} = 9.58$, $P < 10^{-6}$; probability: $\beta = 0.46 \pm 0.04$, $t_{17} = 10.78$, $P < 10^{-6}$). This argues against the possibility that confidence could be a direct readout of reaction time and favors the hypothesis that reaction time is modulated by the same uncertainty as confidence, as well as integrating other sources of variability.

Study 3b: neural representation of confidence in desirability and probability rating

A last group of healthy participants ($n = 26$) was scanned while performing the same task as in study 3a, with the same prospects and judgments, except that no confidence rating was required (**Fig. 6a**). Neural activation was modeled as a boxcar covering both prospect presentation and the rating period, with desirability and probability trials in separate regressors. We included the rating period because, at the time of prospect presentation, subjects were not aware of which type of judgment (desirability or probability) would follow. Boxcars were parametrically modulated by response time, in addition to second-order polynomial extension of ratings (with linear and quadratic terms in separate regressors). We first analyzed the modulation by desirability and probability separately, to replicate the finding that the VMPFC encodes subjective value and not any first-order rating. We indeed observed a significant VMPFC activation with desirability (**Fig. 6b**, $P_{FWE\_clu} < 0.05$) but not with probability, even at a permissive threshold ($P_{unc} < 0.01$). Note that owing to the psychological phenomenon termed optimism bias[22], desirability and probability judgments were correlated ($\rho = 0.30 \pm 0.04$, $t_{25} = 8.25$, $P < 10^{-6}$). However, orthogonalizing the two types of rating did not change any of the fMRI results. If anything, the correlation with desirability should have played against our result regarding probability (no relationship with VMPFC activity).

Then we tested the hypothesis that the VMPFC would also encode confidence with a triple conjunction, including the linear and quadratic extensions of desirability ratings, as well as the quadratic extension of probability rating. At the whole brain level, this conjunction analysis revealed a single cluster, extending over the frontopolar and ventromedial prefrontal cortices (**Fig. 6c**, $P_{FWE\_clu} < 0.05$). These results were confirmed by an analysis focused on our VMPFC ROI, independently defined in study 1b. Regarding desirability judgment, regression estimates extracted from this VMPFC ROI were both highly significant and had a similar effect size for the linear and quadratic functions (one-sample *t* tests on individual regression estimates; linear: $\beta = 0.08 \pm 0.02$, $t_{25} = 4.63$, $P = 9.6 \times 10^{-5}$; quadratic: $\beta = 0.09 \pm 0.02$, $t_{25} = 4.56$, $P = 1.2 \times 10^{-4}$). In contrast, regression estimates obtained for probability judgment did not follow a linear function, but only a quadratic one (linear: $\beta = 0.02 \pm 0.01$, $t_{25} = 1.54$, $P = 0.14$; quadratic: $\beta = 0.06 \pm 0.02$, $t_{25} = 3.02$, $P = 0.0058$). Again these results were similar whether or not response time was included in the regression model.

To illustrate encoding of value and confidence in the VMPFC, signal change was plotted against nine bins of ascending desirability or probability (**Fig. 6d**) and effect size obtained with a parametric FIR model was plotted as a function of time (**Fig. 6e**). Study 3b confirmed the notion that the VMPFC encodes specifically the first-order judgments related to value, and not belief. The U-shape

function of probability observed in VMPFC activity further suggests that it encodes a second-order confidence judgment, rather than stimulus salience.

DISCUSSION

A highly replicated finding in neuroeconomics is that value estimates are encoded in a set of specific brain regions, among which the VMPFC appears to be key[3–5]. Here we demonstrate that VMPFC activity is not a pure reflection of stimulus value because it also incorporates confidence, understood as a second-order judgment on the correctness of a first-order response. Furthermore, we suggest that the integration of response confidence is a general and automatic phenomenon. In the following we successively discuss the computational, behavioral and neuroimaging aspects of our demonstration.

Our computational model formalizes the notion that confidence is a second-order judgment on the correctness of a first-order response. This definition follows on a long tradition in the psychophysics of perceptual decisions[23–25]. In these early studies, as in more recent accounts[17,18,26,27], confidence has been conceived as a secondary representation arising from the amount of perceptual evidence on which decisions are based. The amount of evidence could be captured by a probability distribution over a decision variable defined as the difference between options in the relevant dimension: for example, a difference in brightness between two visual stimuli. We intended to generalize this idea in several ways. First, the model applies to external signals that can be objectively measured (for example, brightness or age judgments) and also to internal signals that are more subject specific (for example, desirability or probability estimates). Second, the model extends to any kind of judgment, including those that involve just one item (for example, a face), instead of restricting the theory to binary choice—that is, to judgment on the difference between two items. Third, the model generalizes to the cases where more than two responses can be given, as implemented in rating tasks (for example, from −10 to 10, not just yes versus no).

Previous studies have derived confidence in binary choice from the endpoint of a decision variable that accumulates the results of sequentially sampling a probability distribution[15,17,18]. We tried to adapt this formalism to rating tasks, but it led to major inconsistencies, such as confidence increasing (not decreasing) with the variance of the probability distribution (see Online Methods and ref. 28). The model that we suggest instead captures the processes involved by rating tasks in a more direct manner. It starts with the intuitive assumption that subjects intend to minimize the gap between their overt rating and their internal judgment. This involves specifying a utility function based on this distance and a mapping function for the projection from internal judgment to external rating. We used standard, parsimonious functions, with a squared distance for the utility function and a sigmoid projection for the mapping function. Confidence was naturally defined as the expected response accuracy: that is, the quantity that subjects intend to maximize. Note there is a slight twist in the notion of accuracy here, as it refers in our model to the adequacy of the external response with respect to a subjective judgment (and not to an objective measure), which allows addressing tasks in which there is no good or bad answer. Under these definitions, analytical decomposition allowed us to derive how rating and confidence should vary with the mean and variance of the internal probability distribution and, most importantly, how they relate to each other.

This formalism captures two intuitions. The first intuition is that we prefer to be more confident, as confidence is the quantity that the model maximizes when selecting a particular rating. This makes a link between confidence and the notion of expected value employed in economic choice. The second intuition is that midscale ratings can be of two sorts: one would be "I know this is indifferent to me" (mean around zero) and the other "I don't know how I feel about this" (high variance). In other words, uncertainty tends to bias the response toward the middle of the scale, which is expressed by the quadratic relationship between rating and confidence.

To our knowledge, the quadratic link between second-order confidence and first-order ratings has never been specifically explored, despite its intuitive appeal. Our behavioral experiments confirmed this quadratic relationship in four types of first-order rating, including subjective dimensions (pleasantness, desirability and probability rating) as well as objective features of the stimuli (age rating). Although the same quadratic pattern was observed on average, we noted some variations related to individual expertise. For instance, in the age rating task, many subjects found modern paintings easier to date than old ones. Individual expertise may therefore induce a departure from a pure quadratic relationship with objective dimensions, for which the idiosyncratic uncertainty is likely to vary across the scale in an asymmetric manner.

The use of age and probability as controls was furthermore important to rule out a possible confound of salience with confidence. Salience is a loosely defined concept; it usually refers to the property of an object that attracts attention. In neuroeconomics, it has been defined as unsigned value, such that very appetitive and very aversive items are both highly salient[29–31]. This means that salience should vary as a V-shaped function of value, which can easily turn into a U shape with the noise inherent in experimental data acquisition. Yet this would only occur if value ratings are centered on the reference point indicating the transition between aversive and appetitive items. However, the observed U-shaped relationship between second-order confidence levels and first-order ratings did not depend on whether we used only appetitive or both appetitive and aversive items, and neither did it depend on whether the rating scale had only positive or both positive and negative numbers. This is consistent with the sigmoid mapping implemented in our model, which normalizes the ratings irrespective of the particular scale imposed on participants. In addition, we also observed a U-shaped relationship between confidence and probability ratings, which allowed us to dissociate confidence from salience. Indeed, salience should vary linearly (not quadratically) with probability, as more probable events are by definition less surprising and therefore less salient.

Yet several limitations of the model must be acknowledged. A first limitation is that the model does not account for how participants adjust the sigmoid mapping depending on the set of stimuli and rating scale they are given. Some anchoring effect is likely to occur, as subjects probably need a series of trials to learn the range of values covered by the stimuli. To avoid this, we trained participants on a set of stimuli that were not identical but were similar in mean and variance to the stimuli used during the experiments. Thus we believe that the behavioral results were acquired after participants had adjusted the mapping from internal values to external ratings.

A second limitation is that the model does not describe the dynamics of the rating process, which would be necessary to predict response time. We observed that reaction time also varied as a U-shaped function of first-order ratings, albeit in a noisier manner suggesting that it also includes other sources of variance. We suspect that reaction time is also affected by the processes constructing the internal judgment (for example, pleasantness) and not just by the processes translating the uncertainty of this judgment into an overt rating. Modeling these processes would go far beyond the scope of this study but certainly represents a key challenge for neuroeconomics.

A third limitation is that our model is formalized at a computational level and does not suggest any neural implementation. Further work would be needed to understand whether the projection of the internal probability distribution to the rating scale and the confidence-based selection of optimal rating are just 'as if' mechanisms or whether they are truly implemented in the brain. In the following, we only highlight one neural correlate: the encoding of confidence in VMPFC activity.

The analysis of old and new fMRI studies revealed that VMPFC activity was correlated with the quadratic extension of first-order ratings, which we take as a proxy for confidence. Note that the quadratic pattern of activity was not a trivial reflection of response time, which was also included in the general linear model. We found no evidence for VMPFC activity decreasing with response time, even when the squared rating was removed from the general linear model. This observation rules out the possibility that the VMPFC could represent not confidence but overall task difficulty (or easiness), which might be quantified by response time.

Somewhat ironically, the VMPFC ROI was defined as a region encoding stimulus pleasantness—that is, correlating linearly with first-order ratings. The meta-analysis using the same ROI across studies confirmed the robustness of value encoding in the VMPFC, despite differences in stimuli (perceived items including faces, houses, paintings, toys, food and clothes, or anticipated events including sport, culture, politics and so forth), instructions (pleasantness or desirability rating) and presentation modes (pictures, videos or sentences). ROI selection should have biased the result toward a linear correlation, but the quadratic term turned out to be significant as well. This means that confidence is encoded in the brain region that has been thought to be the location of subjective valuation[9], as was previously shown in the case of economic choice[15].

Such a finding may not be surprising under our model, in which confidence is construed as the intrinsic value of the behavioral response. However, as in any model-based fMRI analysis, the correlation with VMPFC activity does not prove that the brain actually uses this confidence signal to guide the response. It could well be that confidence is an epiphenomenon—that is, a by-product of a response that is based on other types of signals. Yet since subjects likely intend to be accurate, it would seem bizarre if they did not use this estimate of response accuracy that is represented in their VMPFC.

In whole-brain searches, we did not find any other regions showing consistent conjunction of value and confidence encoding across studies. Like any null result, this absence of effect must be taken with caution. While we are conclusive about the VMPFC, we remain open to the possibility that other regions of the so-called brain valuation system also incorporate confidence. This has been suggested in the literature investigating confidence in recognition memory, which has been related to activity in the ventral striatum and posterior cingulate cortex, in addition to the VMPFC[32–34]. Also, a recent meta-analysis has shown that the most consistent regions encoding confidence judgment across memory and perception tasks are the VMPFC, hippocampus and posterior cingulate cortex[35].

Because the VMPFC exhibited the same quadratic pattern of activity in three fMRI data sets across a variety of first-order ratings (age, pleasantness, probability, desirability), we believe this is a general and robust phenomenon. However, one may find artificial the dissociation of the activity pattern observed with value-related ratings (pleasantness and desirability) into linear and quadratic terms. The notion that two variables (linear, value; quadratic, confidence) were encoded in these conditions is supported by the fact that the quadratic shape was isolated when using first-order ratings that have no value component (age and probability). The latter result also offers evidence that the VMPFC represents only value-related judgment, and not any type of rating. A U-shaped pattern has already been noted in the VMPFC or other brain valuation system regions and has been interpreted as reflecting arousal or salience[29,36–38]. A more parsimonious interpretation that accounts for all our results is that the VMPFC encodes confidence (or response value) in addition to stimulus value. The idea of a common neural currency has been separately suggested for value[39] and confidence[40]. Our findings further suggest that the currency represented by the VMPFC signal might be common to value and confidence.

A crucial finding is that confidence encoding in the VMPFC was observed in the absence of any instruction to report confidence. This is consistent with the subjective experience that a feeling of confidence (or doubt) automatically arises when making a judgment. Our model provides a reason for this automaticity: it comes from confidence being the quantity that is maximized (and therefore estimated) when selecting the best response. Thus, the automatic integration of confidence demonstrated here strengthens the general view of the VMPFC as an automatic valuation system, which has been previously documented in the case of stimulus values[9,11–13].

We believe that our findings might bring important constraints for building a biologically inspired theory of choice. A first constraint is that any valuation operated by the VMPFC during choice or rating tasks might come with a confidence level (or an uncertainty level). The uncertainty that is typically considered in economic choice is the stochastic component of drift diffusion models: that is, an uncertainty attached to the value difference between the two options. Taking into account the confidence that subjects have in each option valuation might affect choice and response time in nontrivial ways. A second constraint is that VMPFC activity might aggregate the value of many aspects of a given situation. We suspect that such an aggregation of values, including response confidence, has been engaged but overlooked in previous studies. Taking this into consideration would complicate decoding the value of one particular item, since this value might be confounded by the values of contextual features, such as task pleasantness or physical discomfort, which could vary across time or conditions. However, unraveling such mechanisms of value aggregation in the VMPFC might be key to explaining many irrational behaviors. Indeed, aggregation of value and confidence might lead to misattribution, such that one could feel more confident not because one expects to perform better but because one is in a more pleasant context.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The study was funded by a Starting Grant for the European Research Council (ERC-BioMotiv) and a Research Grant from the Schlumberger Foundation. M.L. received a PhD fellowship from the French Ministère de la Recherche and an Amsterdam Brain and Cognition Talent Grant from the University of Amsterdam. R.A. received a PhD fellowship from the Direction Générale de l'Armement and a grant from the Fondation pour la Recherche Médicale. This work also benefited from the program "Investissements d'avenir" (ANR-10-IAIHU-06). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

M.L. and M.P. designed all experiments. M.L. and R.A. collected the data. M.L. performed the data analysis. J.D. formalized the computational model. M.L. and M.P. wrote the manuscript. All authors discussed the results and commented the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Von Neumann, J. & Morgenstern, O. *Game Theory and Economic Behavior* (Princeton Univ. Press, 1944).
2. Samuelson, P.A. A note on the pure theory of consumer's behaviour. *Economica* **5**, 61–71 (1938).
3. Peters, J. & Buchel, C. Neural representations of subjective reward value. *Behav. Brain Res.* **213**, 135–141 (2010).
4. Bartra, O., McGuire, J.T. & Kable, J.W. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).
5. Clithero, J.A. & Rangel, A. Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* **9**, 1289–1302 (2014).
6. Blood, A.J., Zatorre, R., Bermudez, P. & Evans, A. Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. *Nat. Neurosci.* **2**, 382–387 (1999).
7. Chib, V.S., Rangel, A., Shimojo, S. & O'Doherty, J.P. Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *J. Neurosci.* **29**, 12315–12320 (2009).
8. Hare, T.A., Camerer, C.F., Knoepfle, D.T., O'Doherty, J.P. & Rangel, A. Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *J. Neurosci.* **30**, 583–590 (2010).
9. Lebreton, M., Jorge, S., Michel, V., Thirion, B. & Pessiglione, M. An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron* **64**, 431–439 (2009).
10. Plassmann, H., O'Doherty, J. & Rangel, A. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J. Neurosci.* **27**, 9984–9988 (2007).
11. Levy, I., Lazzaro, S., Rutledge, R. & Glimcher, P. Choice from non-choice: predicting consumer preferences from blood oxygenation level-dependent signals obtained during passive viewing. *J. Neurosci.* **31**, 118–125 (2011).
12. Harvey, A.H., Kirk, U., Denfield, G. & Montague, P. Monetary favors and their influence on neural responses and revealed preference. *J. Neurosci.* **30**, 9597–9602 (2010).
13. Abitbol, R. *et al.* Neural mechanisms underlying contextual dependency of subjective values: converging evidence from monkeys and humans. *J. Neurosci.* **35**, 2308–2320 (2015).
14. Kepecs, A., Uchida, N., Zariwala, H. & Mainen, Z. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
15. De Martino, B., Fleming, S.M., Garrett, N. & Dolan, R.J. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110 (2013).
16. Rolls, E.T., Grabenhorst, F. & Deco, G. Choice, difficulty, and confidence in the brain. *Neuroimage* **53**, 694–706 (2010).
17. Pleskac, T.J. & Busemeyer, J.R. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. Rev.* **117**, 864–901 (2010).
18. Yu, S., Pleskac, T.J. & Zeigenfuse, M.D. Dynamics of postdecisional processing of confidence. *J. Exp. Psychol. Gen.* **144**, 489–510 (2015).
19. Griffin, D. & Tversky, A. The weighing of evidence and the determinants of confidence. *Cognit. Psychol.* **24**, 411–435 (1992).
20. Lichtenstein, S., Fischhoff, B. & Phillips, L.D. in *Heuristics and Biases* 306–334 (Cambridge Univ. Press, 1982).
21. Lebreton, M., Kawa, S., Forgeot d'Arc, B., Daunizeau, J. & Pessiglione, M. Your goal is mine: unraveling mimetic desires in the human brain. *J. Neurosci.* **32**, 7146–7157 (2012).
22. Sharot, T., Riccardi, A.M., Raio, C.M. & Phelps, E.A. Neural mechanisms mediating optimism bias. *Nature* **450**, 102–105 (2007).
23. Pierce, C.S. & Jastrow, J. On small differences of sensation. *Mem. Natl. Acad. Sci.* **3**, 73–83 (1884).
24. Adams, J.K. A confidence scale defined in terms of expected percentages. *Am. J. Psychol.* **70**, 432–436 (1957).
25. Vickers, D. *Decision Processes in Visual Perception* (Academic, New York, 1979).
26. Fleming, S.M. & Dolan, R.J. The neural basis of metacognitive ability. *Phil. Trans. R. Soc. Lond. B* **367**, 1338–1349 (2012).
27. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Phil. Trans. R. Soc. Lond. B* **367**, 1310–1321 (2012).
28. Daunizeau, J. A note on race models http://sites.google.com/site/jeandaunizeauswebsite/links/resources (2015).
29. Litt, A., Plassmann, H., Shiv, B. & Rangel, A. Dissociating valuation and saliency signals during decision-making. *Cereb. Cortex* **21**, 95–102 (2011).
30. Maunsell, J.H. Neuronal representations of cognitive state: reward or attention? *Trends Cogn. Sci.* **8**, 261–265 (2004).
31. Roesch, M.R. & Olson, C.R. Neuronal activity related to anticipated reward in frontal cortex: does it represent value or reflect motivation? *Ann. NY Acad. Sci.* **1121**, 431–446 (2007).
32. Chua, E.F., Schacter, D.L., Rand-Giovannetti, E. & Sperling, R.A. Understanding metamemory: neural correlates of the cognitive process and subjective level of confidence in recognition memory. *Neuroimage* **29**, 1150–1160 (2006).
33. Moritz, S., Glascher, J., Sommer, T., Buchel, C. & Braus, D.F. Neural correlates of memory confidence. *Neuroimage* **33**, 1188–1193 (2006).
34. Schwarze, U., Bingel, U., Badre, D. & Sommer, T. Ventral striatal activity correlates with memory confidence for old- and new-responses in a difficult recognition test. *PLoS ONE* **8**, e54324 (2013).
35. White, T.P., Engen, N.H., Sørensen, S., Overgaard, M. & Shergill, S.S. Uncertainty and confidence from the triple-network perspective: voxel-based meta-analyses. *Brain Cogn.* **85**, 191–200 (2014).
36. Costa, V.D., Lang, P.J., Sabatinelli, D., Versace, F. & Bradley, M.M. Emotional imagery: assessing pleasure and arousal in the brain's reward circuitry. *Hum. Brain Mapp.* **31**, 1446–1457 (2010).
37. Elliott, R., Newman, J.L., Longe, O.A. & Deakin, J.F.W. Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: a parametric functional magnetic resonance imaging study. *J. Neurosci.* **23**, 303–307 (2003).
38. Cooper, J.C. & Knutson, B. Valence and salience contribute to nucleus accumbens activation. *Neuroimage* **39**, 538–547 (2008).
39. Levy, D. & Glimcher, P.W. The root of all value: a neural common currency for choice. *Curr. Opin. Neurosci.* **22**, 1027–1038 (2012).
40. de Gardelle, V. & Mamassian, P. Does confidence use a common currency across two visual tasks? *Psychol. Sci.* **25**, 1286–1288 (2014).

ONLINE METHODS

Computational model. In what follows, we describe a computational model that attempts to explain two phenomena: (i) confidence is a U-shaped function of first-order judgments, and (ii) confidence is linked to the intrinsic value of the behavioral response. We will consider a pleasantness rating task similar to that used in study 1.

Let x measures how pleasant the item to be rated is on people's internal (subjective) scale. People are asked to rate the item pleasantness on a bounded arbitrary scale. Without loss of generality, we assume that the external scale is bounded between 0 and 1. This implies that people need to map their internal (natural) pleasantness judgment x onto the [0,1] interval. Note that to preserve preference orderings, this mapping must be monotonic. Again, without loss of generality, the sigmoid mapping $s{:}x \rightarrow s(x) = 1/(1+e^{-x})$ is the simplest (smoothest) mapping that satisfies these constraints. Let r be the participant's rating on the external scale. The task instructions can now be understood as follows: find a rating $\hat{r}$ that best matches the mapped pleasantness $s(x)$. This induces a utility function $U(r,x) = -[r - s(x)]^2$ that measures the accuracy of the mapped pleasantness. Here the accuracy is simply the negative of the squared error of the rating. Complying with the task instructions thus reduces to maximizing the utility function $U(r,x)$; that is, maximizing the accuracy of the rating.

If x was known without ambiguity, the solution to this problem would be trivial:

$$\begin{aligned} \hat{r} &= \arg\max_r U(r,x) \\ &= s(x) \end{aligned} \tag{1}$$

Equation (1) means that, in the absence of uncertainty about x, the optimal rating $\hat{r}$ is simply $s(x)$.

Now let us assume that people are uncertain about x: that is, the information people have about x is captured by a probability distribution function $p(x)$ with mean μ and variance σ. This means that the utility $U(r,x)$ cannot be evaluated (it is itself uncertain). The decision-theoretic solution to the rating task is thus to maximize the expected utility, which yields a rating that is on average optimal (**Fig. 1a**). In other words, the optimal rating $\hat{r}$ is the solution to the following problem:

$$\begin{aligned} \hat{r} &= \arg\max_r Q(r,\mu,\sigma) \\ Q(r,\mu,\sigma) &= E\left[U(r,x)\right] \end{aligned} \tag{2}$$

where $Q(r, \mu, \sigma)$ is the expected utility (which depends upon the moments of $p(x)$) and the expectation is taken under the probability distribution $p(x)$. This quantity is important because confidence is, by definition, the expected accuracy of the rating.

One can show that the expected accuracy can be broken down into two terms:

$$Q(r,\mu,\sigma) = -\left(r - E\left[s(x)\right]\right)^2 - V\left[s(x)\right] \tag{3}$$

where $V[s(x)]$ is the variance of $s(x)$ under $p(x)$.

Inserting equation (3) into equation (2) yields optimal rating $\hat{r}$:

$$\begin{aligned} \hat{r} &= \arg\min_r \left(r - E\left[s(x)\right]\right)^2 \\ &= E\left[s(x)\right] \\ &\approx s\left(\frac{\mu}{\sqrt{1+\sigma\frac{3}{\pi^2}}}\right) \end{aligned} \tag{4}$$

where the first line derives from noting that the first term in equation (3) is the only one that explicitly depends on r and the last line is an analytical approximation to the expected sigmoid mapping[41]. Equation (4) tells us how the optimal rating $\hat{r}$ depends on μ and σ, which measures how uncertain people are about the item's pleasantness. **Figure 1b** (left panel) shows how $E[s(x)]$ varies as a function of μ and σ.

In particular, one can show that $\hat{r} \xrightarrow{\sigma\to\infty} 1/2$; that is, people should aim for the middle of the rating scale when they are maximally uncertain. This explains why items that are rated at the middle of the external scale consist of a mixture of items that are certainly neutral (for example, point P_1 in **Fig. 1b**) and items whose pleasantness is uncertain (for example, point P_2 in **Fig. 1b**).

Now equation (3) tells us something more: having rated items optimally, people are left with some residual (nonzero) expected error. This means that, although people may optimally rate the item on the external pleasantness scale, they still expect to be wrong. In turn, optimal confidence measures how accurate people expect to be, having chosen the optimal rating:

$$\begin{aligned} Q(\hat{r},\mu,\sigma) &= -V\left[s(x)\right] \\ &\approx \hat{r}(\hat{r}-1)\left(1 - \frac{1}{\sqrt{1+\sigma\frac{3}{\pi^2}}}\right) \end{aligned} \tag{5}$$

where the first line derives from inserting equation (4) into equation (3) and the second line is an analytical approximation to the variance of the sigmoid mapping[41]. **Figure 1b** (right panel) shows how $V[s(x)]$ varies as a function of μ and σ. One can show that $\delta Q/\delta\sigma \leq 0$; that is, optimal confidence is a monotonically decreasing function of people's uncertainty σ (see **Fig. 1c**, left). Besides, $Q(\hat{r},\mu,\sigma) \xrightarrow{\sigma\to 0} 0$; that is, there is no residual error when people are certain about how pleasant the item is.

Taken together, equations (4) and (5) make two predictions: (i) confidence $\hat{q} = Q(\hat{r},\mu,\sigma)$ is a quadratic function of rating $\hat{r}$ (see **Fig. 1c**, right panel), and (ii) being confident about the rating has high utility (in terms of people's feeling about being accurate). The latter provides a normative explanation for why being confident about the rating is, in and of itself, valuable.

Lastly, note that the above two predictions hold irrespective of the nature of the rating, as long as the internal uncertainty σ is non-negligible when compared to the extent of the external rating scale.

We considered alternative models for our empirical findings. For example, De Martino and colleagues[15] propose a model for confidence judgments in the context of binary decisions, which can be extended to pleasantness rating tasks as follows. One would assume that two accumulators representing the left ("I do not like it") and right ("I like it") halves of the rating scale enter a race that ends whenever one of the accumulator reaches a predefined threshold. The winning accumulator determines both the rating time and the rating sign (if the central position is zero). Both rating magnitude and confidence would then be given by the distance between the two accumulators. Analysis of this race model reveals a nontrivial consequence of the above definition of confidence: confidence level should increase, not decrease, with the variance of the value signal. This contradiction drove us to conclude that reading confidence as the endpoint of an accumulation process is not appropriate, at least in the case of rating tasks. Most relevant mathematical details are laid out in a technical note on race models that is available online[28].

Subjects. The studies were approved by the Ethics Committee for Biomedical Research of the Pitié-Salpêtrière Hospital, where they were conducted. Subjects were recruited via the RISC (Relais d'Information en Sciences Cognitives) website and screened for exclusion criteria: age below 18 or above 39, regular use of drugs or medications, history of psychiatric or neurological disorders and contraindications to MRI scanning (pregnancy, claustrophobia, metallic implants). All subjects gave informed consent before taking part in the study. Unless otherwise specified, subjects were paid 30 for the simple behavioral experiments (studies 1a and 3a) and 100 for the fMRI experiments (studies 1b, 2 and 3b).

A total of 125 subjects were included in the different studies (study 1a, first group: n = 19, 9 males, age = 22.2 ± 2.4; study 1a, second group: n = 22, 12 males, age = 24.4 ± 2.9; study 1b: n = 20, 10 males, age = 22.0 ± 2.7; study 2: n = 19, 11 males, age = 23.9 ± 4.0; study 3a: n = 19, 10 males, age = 23.1 ± 5.3; study 3b: n = 26, 12 males, age 25.3 ± 5.5). Three subjects were excluded, one in study 1a for abnormally long reaction times (>10 s), one in study 2 because

doi:10.1038/nn.4064

a technical problem occurred during scanning, and the last in study 3a for always giving the same confidence level.

Tasks. All tasks were programmed on a PC using the Cogent 2000 (Wellcome Department of Imaging Neuroscience, London, UK) library of Matlab functions for stimulus presentation. They all involved rating procedures that were implemented as follows.

In behavioral studies, subjects performed the task on a standard computer. They could move the cursor by pressing left and right arrows on the keyboard. Ratings were all self-paced, and subjects had to press the spacebar to validate their response and go to the next trial.

In fMRI studies, subjects could move the cursor by pressing a button with the right index finger to go left or with the right middle finger to go right. Again, ratings were all self-paced, and subjects had to press a button with the left index finger to validate their response and go to the next trial.

In both fMRI and behavioral studies, the initial position of the cursor on the scale was randomized to avoid confounding the ratings with the movements they involved.

Details specific to the different tasks are described below.

Study 1. Stimuli were 120 faces, 120 houses and 120 paintings, for a total of 360 pictures that we had used in a previous experiment (see picture selection in ref. 9). The 360 stimuli were randomly distributed over 6 sessions of 60 trials each (20 faces, 20 houses and 20 pictures).

Study 1a. For the first group, every trial started with a fixation cross, after which one picture was displayed on the center of the screen, at the top of a 21-step rating scale graduated between −10 and 10. Participants were asked to indicate on this scale how pleasant the presented stimulus was. After validation of the pleasantness judgment, a sentence reminding participants of their rating appeared ("You gave a rating of *X*"), together with another 100-step (almost continuous) rating scale, on which they were asked to indicate how confident they were about their first-order rating ("How confident are you?", between "Not at all", and "Totally"). The task was similar for the second group, except that subjects had to rate how old (instead of how pleasant) the presented stimulus was, on a 21-step scale that was adapted to the category (face, house or painting).

Study 1b. This fMRI study is a reanalysis of data obtained in a previous experiment (see ref. 9 for detailed methods). Subjects performed three sessions of the pleasantness rating task and three sessions of the age rating tasks, the order being counterbalanced across subjects. In every trial, the picture was first displayed on the screen for 3 s, following a fixation cross. Then the rating scale appeared, and participants had to indicate on this scale how pleasant or how old the presented stimulus was. There was no confidence rating in this study.

Study 2. This fMRI study is a reanalysis of data obtained in a previous experiment (see ref. 21 for detailed methods) using a desirability rating task. Stimuli were 240 short (2–5 s) videos featuring different objects (food, toys, clothes and tools), randomly distributed over four 60-trial sessions. In every trial, the video was first played on the screen, following a fixation cross. Then a 0–10 rating scale appeared, and participants had to indicate how much they would like to acquire the object. There was no confidence rating in this study.

Study 3. Stimuli were 270 potential events (prospects; see examples in **Supplementary Table 1**) from various domains (politics, sport, society, culture, media, economics, diplomacy, science, technology, etc.). Subjects were instructed to read the text depicting the event and think of how pleased they would feel should this event happen in the next 5 years (for desirability rating) and how likely they estimated this event would be to actually happen in the next 5 years (for probability rating). They were randomly distributed over 5 sessions of 54 trials each (27 desirability rating and 27 probability rating trials, randomly intermixed).

Study 3a. Every trial started with a fixation cross, after which the event was displayed on the screen for a duration of 2 s. Then appeared a 21-step rating scale that could be either a desirability or a probability scale. The desirability scale was graduated between −10 (not desirable at all) and 10 (highly desirable), whereas the probability scale was graduated from 0 (completely unlikely) to 100% (most likely). The scales were accompanied by the word "DESIRABILITY" or "PROBABILITY," which served as instruction. After validation of the first-order rating, a continuous scale was displayed, on which the subject had to indicate how confident they were ("How confident are you?", between "Not at all" and "Totally") about their first-order rating, which was recalled on screen ("You gave a rating of *X*").

Study 3b. The preceding task was adapted to comply with fMRI limitations, as well as to assess the automaticity of confidence elicitation. On every trial, one prospect was displayed alone on the screen for a duration drawn from a uniform 5–7 s distribution, following a 1-s fixation cross. The desirability or probability scale only appeared after prospect display, and there was no confidence rating. We also added an extra jitter: one out of eight trials started with a fixation cross lasting 9 s instead of 1 s. Outside the scanner, participants underwent five other sessions where each event was presented a second time, and participants were asked to rate the dimension complementary to the one that was randomly selected in the scanner (that is, probability judgment for events that were assessed on the value dimension in the scanner, and vice versa).

Statistics. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those generally employed in the field. Unless otherwise specified, all dependent variables (raw, *z*-scored or binned behavioral measures and robust regression estimates) were computed at the session level, averaged at the subject level and tested for significance at the group level (random effect analysis) using two-tailed paired *t*-tests. All robust regressions were performed on *z*-scored independent and dependent variables. Data distribution was assumed to be normal but this was not formally tested. All statistical analyses were performed with Matlab Statistical Toolbox (Matlab R2014a, The MathWorks, Inc., USA).

Code availability. The code used to generate the simulations illustrating our theoretical model in **Figure 1** and to run the behavioral tasks is available upon request.

Neuroimaging data acquisition. For all imaging studies, T2*-weighted echo planar images (EPI) were acquired with blood oxygen level–dependent (BOLD) contrast on a 3.0-T magnetic resonance scanner. We employed a tilted plane acquisition sequence designed to optimize functional sensitivity in the orbitofrontal cortex and medial temporal lobes[42]. To cover the whole brain with good spatial resolution, we used the following parameters:

Study 1b: TR = 2.29 s, 35 slices, 2 mm slice thickness, 1 mm inter-slice gap
Study 2: TR = 2.0 s, 35 slices, 2 mm slice thickness, 1.5 mm inter-slice gap.
Study 3b: TR = 2.03 s, 35 slices, 2 mm slice thickness, 1.6 inter-slice gap.

T1-weighted structural images were also acquired, co-registered with the mean EPI, normalized using nonlinear transformation to a standard T1 template, and averaged across subjects to allow group level anatomical localization. EPI data were analyzed in an event-related manner, within a general linear model, using the statistical parametric mapping software SPM8 (Wellcome Trust center for NeuroImaging, London, UK) implemented in Matlab. The first 5 volumes of each session were discarded to allow for T1 equilibration effects. Preprocessing consisted of spatial realignment, normalization using the same transformation as structural images, and spatial smoothing using a Gaussian kernel with a full-width at half-maximum (FWHM) of 8 mm.

Neuroimaging data analysis. *General linear models (GLM).* We used the following GLM to explain subject-level time series.

Study 1a. Events were image display, modeled as boxcar function. This categorical regressor was modulated by two parameters accounting for the first- and second-order polynomial expansion of ratings (either age or pleasantness, depending on the session), which were *z*-scored per category (face, house, painting) beforehand. We also modeled the rating period in another regressor with a boxcar function modulated by response time.

Study 2. Events were video display, modeled as boxcar function. This categorical regressor was modulated by two parameters accounting for the first- and second-order polynomial expansion of desirability ratings. We also modeled the rating period in another regressor with a boxcar function modulated by response time.

Study 3b. The two types of trial, corresponding to desirability and probability rating, were modeled in separate regressors, as boxcar functions covering both stimulus presentation and rating scale. Those events were modulated by four parameters each: the first- and second-order polynomial expansion of *z*-scored rating (either desirability or probability, depending on trial type), the response time and the prospect length (number of characters).

Whole-brain analysis. All regressors of interest were convolved with a canonical hemodynamic response function. To correct for motion artifacts, subject-specific realignment parameters were modeled as covariates of no interest. Linear contrasts of regression coefficients (β values) were computed at the session level, averaged at the subject level and taken to a group-level random effect analysis, using one-sample *t*-tests. Conjunction analyses[43] were performed to find brain regions encoding two or three parameters of interest.

Unless otherwise specified, all activations maps were thresholded using family-wise correction for multiple comparison (FWE) either at the voxel level ($P_{FWE\_vox} < 0.05$) or at the cluster level ($P_{FWE\_clu} < 0.05$). This cluster-wise correction was estimated by SPM8 using cluster-generating voxel-level thresholds of $P_{unc\_vox} < 0.001$ or $P_{unc\_vox} < 0.005$, which led to a minimum cluster size of $k \approx 150–250$ voxels (depending on the map being considered).

Region of interest (ROI). The region used in all ROI analyses is the group-level significant cluster revealed in **Figure 3b** (top), which corresponds to the brain region that linearly encoded pleasantness rating at a very stringent statistical threshold ($P_{FWE\_vox\ p} < 0.05$ and a minimal cluster size of $k = 100$ voxels) in study 1b.

Bin models. To examine the relationship between the BOLD signal in our ROI and parameters of interest, we constructed bin models. Trials were ranked according to the parameter of interest (for example, pleasantness or desirability rating) and sorted into nine or ten bins (depending on the total number of trials). Trials corresponding to different bins were modeled in separate regressors of a standard GLM, as boxcar functions over stimulus display (study 1b) or over stimulus display plus the rating period (study 3b). These events were convolved with a canonical hemodynamic response function.

Finite-impulse response (FIR) models. To examine the time course of the relationship between BOLD signal and first- or second-order polynomial expansions of subjective ratings, we built parametric FIR models. For studies 1b and 2b, we modeled eight events as sticks every TR (2 s), starting at the onset of stimuli. Each of these events was modulated by both the first and second polynomial expansions of *z*-scored ratings. For study 3b, desirability and likelihood trials were modeled separately, with sticks every TR (2 s) starting at the stimulus onset (sentence display), modulated by the linear expansions of *z*-scored ratings and a parameter of no interest accounting for reading time (number of letters in the sentence). We also modeled 11 sticks every TR (2 s) locked on the rating scale onset, from which we subtracted the average duration of the stimulus (6 s), such that these events started on average at the same time as in the previous FIR model. These last 11 events were modulated by the second expansion of *z*-scored ratings and a parameter of no interest accounting for stimulus duration. To correct for motion artifacts, subject-specific realignment parameters were modeled as covariates of no interest. Regression estimates were computed at the session level, extracted from our ROI, averaged at the subject level and then plotted at a group level.

A **Supplementary Methods Checklist** is available.

41. Daunizeau, J. On the exponential, sigmoid and softmax mappings http://sites.google.com/site/jeandaunizeauswebsite/links/resources (2014).
42. Deichmann, R., Gottfried, J., Hutton, C. & Turner, R. Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage* **19**, 430–441 (2003).
43. Friston, K.J., Penny, W.D. & Glaser, D.E. Conjunction revisited. *Neuroimage* **25**, 661–667 (2005).

doi:10.1038/nn.4064