The role of the neural reward circuitry in self-referential optimistic belief updates

Kuzmanovic, Bojana; Jefferson, Anneli; Vogeley, Kai

DOI:
10.1016/j.neuroimage.2016.02.014

Document Version
Peer reviewed version

Citation for published version (Harvard):
Kuzmanovic, B, Jefferson, A & Vogeley, K 2016, 'The role of the neural reward circuitry in self-referential optimistic belief updates', *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2016.02.014

The role of the neural reward circuitry in self-referential optimistic belief updates

Bojana Kuzmanovic, Anneli Jefferson, Kai Vogeley

PII: S1053-8119(16)00121-X
DOI: doi: 10.1016/j.neuroimage.2016.02.014
Reference: YNIMG 12939

To appear in: *NeuroImage*

Received date: 14 August 2015
Accepted date: 8 February 2016

Please cite this article as: Kuzmanovic, Bojana, Jefferson, Anneli, Vogeley, Kai, The role of the neural reward circuitry in self-referential optimistic belief updates, *NeuroImage* (2016), doi: 10.1016/j.neuroimage.2016.02.014

The role of the neural reward circuitry in self-referential optimistic belief updates

Bojana Kuzmanovic[a,b,c], Anneli Jefferson[d], Kai Vogeley[c,e]

[a] Max Planck Institute for Metabolism Research Cologne, Germany;

[b] Research Center Juelich, Institute of Neuroscience and Medicine, Ethics in the Neurosciences (INM-8), Germany;

[c] University Hospital Cologne, Department of Psychiatry and Psychotherapy, Germany;

[d] University of Birmingham, Department of Philosophy, UK;

[e] Research Center Juelich, Institute of Neuroscience and Medicine, Cognitive Neuroscience (INM-3), Germany.

Corresponding author:

Bojana Kuzmanovic

Mailing address: Max Planck Institute for Metabolism Research,

Gleulerstr. 50, 50931 Cologne, Germany

E-Mail: bojana.kuzmanovic@sf.mpg.de

Phone: +49 221 4725-255

Fax: +49 221 4725-344

Abstract

People are motivated to adopt the most favorable beliefs about their future because positive beliefs are experienced as rewarding. However, it is so far inconclusive whether brain regions known to represent reward values are involved in the generation of optimistically biased belief updates. To address this question, we investigated neural correlates of belief updates that result in relatively better future outlooks, and therefore imply a positive subjective value of the judgment outcome. Participants estimated the probability of experiencing different *adverse* future events. After being provided with population base rates of these events, they had the opportunity to update their initial estimates. Participants made judgments concerning *themselves* or a similar *other*, and were confronted with *desirable* or *undesirable* base rates (i.e. lower or higher than their initial estimates).

Belief updates were smaller following undesirable than desirable information, and this optimism bias was stronger for judgments regarding oneself than others. During updating, the positive value of *self-related updates* was reflected by neural activity in the subgenual ventromedial prefrontal cortex (vmPFC) that increased both with increasing sizes of favorable updates, and with decreasing sizes of unfavorable updates. During the processing of *self-related undesirable base rates*, increasing activity in a network including the dorsomedial PFC, hippocampus, thalamus and ventral striatum predicted decreasing update sizes.

Thus, key regions of the neural reward circuitry contributed to the generation of optimistically biased self-referential belief updates. While the vmPFC tracked subjective values of belief updates, a network including the ventral striatum was involved in neglecting information calling for unfavorable updates.

Keywords: belief update; optimism bias; reward; subjective value; subgenual vmPFC.

1 Introduction

Thinking about the future is part of a person's identity and supports action planning, decision-making and emotion regulation (Carver and Scheier, 2014; D'Argembeau et al., 2012; D'Argembeau et al., 2009). However, this highly influential prospective thinking does not provide us with the most realistic future outlook, but is instead prevalently optimistically biased. Cross-culturally and independently of gender and age, people tend to overestimate the likelihood of positive future outcomes, and to underestimate the likelihood of negative ones in various domains of daily life, including health-related issues, social relations, and professional success (Chowdhury et al., 2014; Leary, 2007; Sharot et al., 2012; Sharot et al., 2011; Shepperd et al., 2002; Shepperd et al., 2013; Weinstein and Klein, 1996). The consequences of optimistically distorted judgments can be positive or negative: Overestimated chances of success may lead to positive feelings and an increase of effort with beneficial effects for the individual and its environment, but may also lead to miscalculations and failures.

It has been assumed that both motivational and cognitive factors contribute to the optimistic bias, and this reciprocal influence has also been more generally described as "motivated cognition" (Hughes and Zaki, 2015). Cognitive explanations focus on how people achieve desired end states of judgments, and refer to selective memory search and conclusions that are biased toward retrieving confirmatory information for rewarding beliefs (Shepperd et al., 2002). Motivational explanations, on the other hand, relate to the pleasure of having favorable beliefs regarding oneself and one's own future, and the resulting desire to adopt such optimistic beliefs (Shepperd et al., 2002). Accordingly, optimism bias has been described as "a motivation to adopt the most rewarding (or least aversive) perspective on future outcomes" (Sharot et al., 2011).

However, previous neuroimaging research on optimism bias reported solely the recruitment of brain regions related to complex cognitive processing, i.e. the inferior and dorsomedial prefrontal cortex (Sharot et al., 2011). Thus, there is still a lack of evidence for

the motivational explanations, which would require a demonstration of recruitment of key structures of the neural reward circuit. While this complex network includes several cortical and subcortical regions, the most prominent structures involved in human value processing are the ventromedial prefrontal cortex (vmPFC) and the ventral striatum (vStr) (Bartra et al., 2013; Clithero and Rangel, 2014; Haber and Knutson, 2010; Kable and Glimcher, 2009; Peters and Buchel, 2010). The vmPFC, and particularly its subgenual part, has robustly been shown to play a critical role in representing the *positive* subjective value of rewards and emotional stimuli (Chase et al., 2015; Levy and Glimcher, 2012). While the vStr has traditionally been related to learning from errors in reward prediction, more recent research supports an integrative view involving both learning and motivational functions (Bartra et al., 2013; Hamid et al., 2016). More specifically, dopaminergic signaling in vStr has been shown to represent values of estimated future rewards, which influence decisions whether to invest effort in actions aiming at these reward states (Hamid et al., 2016).

The aim of the present study was to demonstrate that favorable beliefs recruit the same brain regions known to be associated with external rewards, to support the view that they have internal positive subjective values able to guide judgment and decision processes. We employed a revised version of an fMRI belief update paradigm. The paradigm assesses how people update their initial beliefs about risks of experiencing hazards when they are provided with base rates for these hazards that result in estimation errors (i.e., the difference between subject's first risk estimation and the presented base rates). It could be shown that updating was optimistically biased because it was larger after *desirable* new information (lower risk than initially expected) than after *undesirable* information (higher risk than initially expected) (Sharot et al., 2011). These results contradict formal learning principles as these predict balanced updating in response to errors, independent of the desirability of the new information.

The first revision was to extend the study design in order to differentiate between judgments referring to oneself and others. Second, in order to increase the experimental control and precision, we systematically manipulated the presented base rates, and included both the first and the second estimation (before and after the presentation of the base rate) in one single trial. And third, we modified the analyses to allow the identification of neural regions that track fluctuations of *updates* on a trial-by-trial level (in contrast to tracking *estimation errors* as in Sharot et al., 2011), because belief updates represent the end state of the judgment process and are expected to have a specific subjective value for judging persons.

We hypothesized that the vmPFC would reflect the differential subjective value of varying update sizes that result in better or worse future outlooks relative to participants' initial beliefs, particularly for self-referential judgments. The larger an update towards an unexpectedly *low* average risk, the better is the subject's adjusted future outlook. Thus, *increasing updates* after *favorable* new information are expected to have *increasing positive value* and to be accompanied by an increasing vmPFC activity. In contrast, the greater an update towards an unexpectedly *high* average risk, the worse is the updated subject's future outlook, so that *decreasing updates* in this condition are expected to result in an *increasing positive value* and increasing vmPFC activity.

Furthermore, we explored in which brain regions the activity during the reception of the new information (base rates) predicted subsequent updates. The belief reconstruction initiated at this point may encompass cognitive processes such as memory retrieval or inferences, but may also be modulated by the motivation to adopt the most favorable conclusions (Shepperd et al., 2002), particularly when these are self-relevant and unfavorable (Sharot et al., 2011). Finally, we explored the relationships between the optimism bias in belief updating and trait optimism.

2 Material and Methods

2.1 Participants

A total of 36 right-handed individuals with no reported history of neurological or psychiatric illness were recruited online within the Research Center Juelich, Germany, and participated in the fMRI study. Twelve persons were excluded from analyses. Five persons had excessive head movements [outliers were selected by total movement greater than 3mm or scan-to-scan motion greater than 1.5 mm, as assessed by ArtRepair Software (Mazaika et al., 2005)], probably due to a relatively tight head coil used in the study; the logfiles of one person were overwritten; one person suspected that the base rates were not correct; two persons had insufficient German language skills as they did not know the meaning of a high number of stimulus events (18 and 24 events). The remaining three persons had a mean positive estimation error of less than 7, because of frequent low fist estimates (see S.1). Thus, data from 24 participants were included in the analyses (mean age = 25.13 years, *SD* = 3.89, ranging from 19 to 38; 13 females). All participants were naïve with respect to the specific purpose of the study, gave written informed consent and were paid for their participation. The study was approved by the local ethics committee of the Medical Faculty of the University of Cologne, Germany.

2.2 Stimulus Material

We used 88 short German descriptions of adverse life events as stimuli and included a wide range of events relating to different life domains (e.g., dementia, arthritis, unemployment, or pest infestation in the home, see Kuzmanovic et al., 2015 for the complete list). The assignment of the stimuli to the experimental conditions and the order of trials were randomized anew for each participant. Note that by applying a random assignment of the stimuli to the experimental conditions, event characteristics that have been suggested to modulate the optimism bias (e.g., base rate, event valence, arousal, controllability, personal experience) (Rose et al., 2008; Sharot et al., 2007; Weinstein, 1980, 1987), or general stimulus characteristics (e.g., number of words or letters) were equally distributed across the

experimental conditions, and thus do not constitute confounding variables (see also Kuzmanovic et al., 2015).

2.3 Design and Procedure

In each trial of the update experiment, participants first had to estimate the probability that different adverse events would occur at least once in a lifetime. Next, they were presented with a corresponding base rate for the general population, and were then given the opportunity to adjust their initial estimate (see Figure 1 for illustration and durations of events). The intervals within the trial ("jitter 4 s" in Figure 1) and between the trials randomly varied, resulting in mean durations of 4000ms and 6000ms, respectively (within-trial durations varied between 2875ms and 5125ms, between-trial durations between 4875ms and 7125ms). The successive arrangement of i) the first estimation, ii) the presentation of the base rate and iii) the second estimation (including the display of the initial estimate) within one trial represent a substantial modification of the original paradigm (Sharot et al., 2011) and served the purpose of minimizing confounding memory effects.

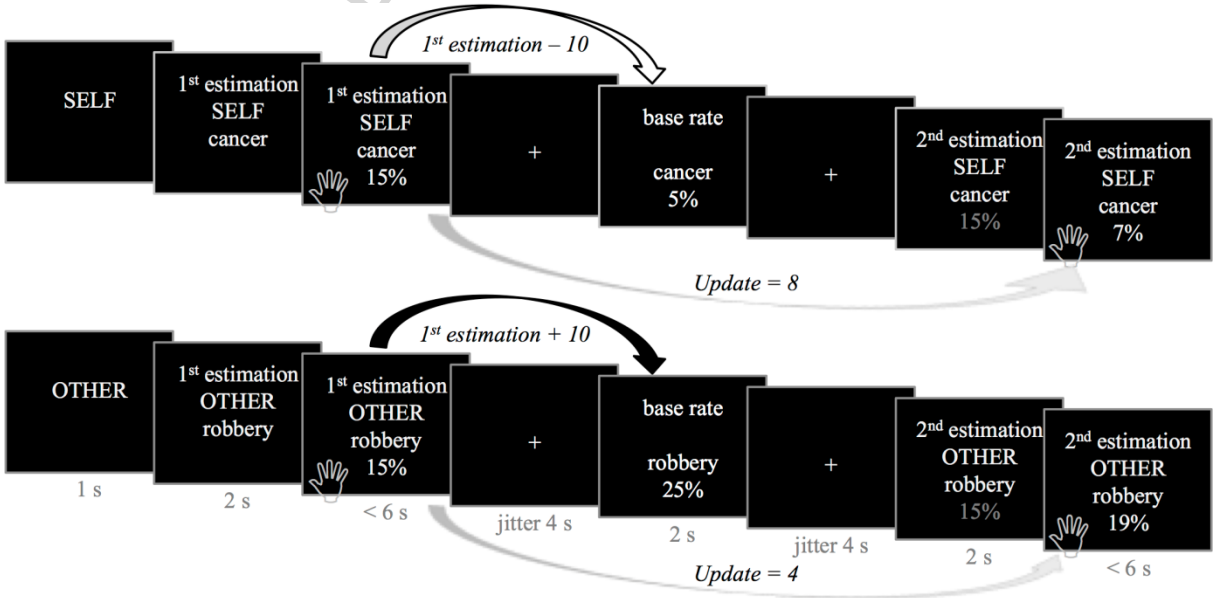

Figure 1 Example trials of the update experiment.
At the beginning of each trial the target person of the judgment was indicated, i.e., self or other. Next, an adverse event was displayed and the participants had to estimate the probability of experiencing it at least once in their (or other's) lifetime. Then, the base rate of the respective event in the general population was presented. Finally, participants had the opportunity to update their initial estimate. Unbeknownst to participants, the valence of the presented base rate was experimentally manipulated by subtracting or adding varying values from participant's first estimate. Upper row: An example of a

desirable (i.e., positive) base rate that was generated by subtracting 10 from the first estimate given. Lower row: An equivalent example of an undesirable (i.e., negative) base rate that was generated by adding 10 to the first estimate given. The exact time course of the trials is shown at the bottom.

The factors that were expected to affect the update behavior within the task were the target person of the judgment (self, other), the valence of the new information (positive, negative), and participants' trait optimism scoring (high, low). In contrast to the original paradigm that included self-related judgments only (Sharot et al., 2011), the target person to whom the probabilities had to be ascribed was manipulated by instructing participants to make estimations for themselves in half of the trials (self), or for a similar other person of the same age, sex and socioeconomic background in the other half of the trials (other). The valence of the new information referring to base rates of *adverse* events depended on participant's initial rating in each trial: When the base rate for an adverse event was *lower* than participants' first estimate, then this constituted desirable or positive (p) information; when it was *higher* than the first estimate, then this constituted undesirable or negative (n) information. Participants' trait optimism was measured by the Life Orientation Test (LOT-R, Glaesmer et al., 2008). The LOT-R median of the whole sample was used to generate two subsamples (median = 17; see Table 2): low trait optimism (low_to, n = 13) and high trait optimism subsample (high_to, n = 11). Thus, a 2-by-2-by-2 design was realized with the factors target (self vs. other) and valence (p vs. n), resulting in four within-subject conditions self_p, self_n, other_p and other_n, and the between-subject factor group (low_to vs. high_to).

Participants were told that the experiment aims to investigate the neural substrates of expectations towards future life events. They were instructed that there was no right or wrong answer as we were interested in their subjective judgment, and to feel free to update their first estimate as much as they wanted. They were also informed that the population base rates were determined by the German Federal Statistical Office ("Statistisches Bundesamt"), and that they should consider this information during their second estimation.

Unbeknownst to participants, the base rates were in fact systematically manipulated in order to control for frequencies and distributions of positive and negative trials, which was the third modification of the original paradigm (Sharot et al., 2011). Positive and negative base rates were computed by subtracting or adding varying values from the first estimate (ranging from 1 to 25; see Figure 1 for an example; see S.1 for more details). In a final debriefing, participants were informed that they had been deceived about the source of population base rates, and the methodological reasons for this procedure were explained.

In addition, participants were instructed to abstain from responding to events for which they felt unable to make a proper estimation. For instance, this could apply to events that participants were experiencing at the time of participation (e.g., currently suffering from hay fever), or that they had never heard about (e.g., unfamiliar diseases).

Responses were given by selecting an absolute probability number with a possible range from 1% to 99% in a completely continuous manner by using three response buttons. Participants always used both hands, in one session the right hand for selecting the percentage number and the left hand for confirming it, and in the other session the other way around (order counter-balanced across participants). The number currently displayed on the screen could be decreased by pressing the left button (e.g., index finger of the right hand), and increased by pressing the right button (e.g., middle finger of the right hand). As soon as the desired number appeared, participants could give a final confirmation of their response by pressing the confirmation button (little finger of the other hand). For all first estimations, the number initially displayed was 50%, and for all second estimations the number corresponded to the selected first estimate (see Figure 1). Participants were instructed to answer within six seconds. If the answer was not recorded within this time, a second and last response period of six seconds was presented. In the event that the first estimate was not provided, the rest of the trial was omitted. Mean durations of the two task sessions were 22.46 min (*SD* = 1.43) and 21.83 min (*SD* = 1.16), respectively.

After the fMRI acquisition, participants completed a short debriefing including ratings of task difficulty, and had the opportunity to describe problems or hypotheses regarding the purpose of the task. Importantly, because we manipulated the base rates, we took great care to assess participants' suspicions regarding their plausibility by using the funnel debriefing method (for more details see Kuzmanovic et al., 2015). Furthermore, they rated their personal experience for each stimulus event that they had seen in the update experiment on a 7-point-scale (from 1 = *unfamiliar* to 7 = *currently experiencing*). The personal experience ratings were carried out on a laptop with the software Presentation (Neurobehavioral Systems, Inc., Version 15.1). Prior to both the update experiment and the personal experience ratings, participants underwent a standardized, computerized instruction including practice trials with stimuli not used within the experimental tasks. Finally, self-report questionnaires (German versions) were completed to measure relevant characteristics of the sample: LOT-R assessing trait optimism on a scale from 0 (pessimistic) to 24 (optimistic), and Beck Depression Inventory (BDI, Beck and Steer, 1987) assessing symptoms of depression on a scale from 0 (minimal) to 63 (maximal).

The MRI data were acquired by using a Magnetom Trio 3T whole body scanner and a 32-channel head coil (Siemens AG, Medical Solutions, Erlangen, Germany). fMRI data during the update experiment were acquired in two sessions with a T2*-weighted gradient echo planar imaging sequence [TR = 2200 ms, TE = 30 ms, field of view = 210 × 210 mm^2, voxel size = 3 × 3 × 3 mm^3, 33 oblique (maximal 30°) axial slices]. Three additional volumes were collected and discarded at the beginning of each session to allow for magnetic saturation. In addition, we acquired high-resolution T1- weighted MPRAGE images (TR = 2250 ms, TE = 3.93 ms, field of view = 256 × 256 mm^2, voxel size = 1 × 1 × 1 mm^3, 128 sagittal images), as well as DTI images (related analyses not reported here). Stimuli and response displays were presented and recorded by the software package Presentation (Neurobehavioral Systems, Inc., Version 15.1), and projected onto a screen at the end of the

magnet bore that participants viewed via a mirror mounted on the head coil. Responses were assessed using a MR-compatible LUMItouch keypad (Photon Control Inc., Burnaby, BC, Canada).

2.4 Measures

The critical measure of participants' behavior was the size of updates after being confronted with base rates of future life events. For each participant, the difference between the first and the second estimate was computed in each of the 88 trials, and then mean updates were computed separately for each of the four experimental conditions (self_p, self_n, other_p and other_n) by averaging all trial-based updates within each condition. Thus, every participant ended up with four repeated-measures of mean updates. If present, trials with missing responses (M = 1.38, SD = 2.00) and trials with estimation error of zero (M = 3.38, SD = 4.20; e.g., when the participant's first estimate was 1%, no errors can be generated to provide desirable base rates, i.e., rates lower than the first estimate; see S.1 for more details) were excluded before computing the mean updates. We used signed update values, but with a differential procedure for lower (positive) and higher (negative) base rates. In conditions self_p and other_p, updates in each trial were computed as first estimate minus second estimate, because the presented base rate was lower than the first estimate and second estimates were expected to be adjusted to this smaller value. Conversely, in conditions self_n and other_n, updates were computed as second estimate minus first estimate. Thus, in the majority of trials independent of condition, updates were equal to or greater than zero (see also Table 1 and Figure 2 for mean values; mean number of zero updates = 20.29, SD = 10.34). For trials in which participants responded in an unexpected direction (e.g., first estimate = 20%, base rate = 10%, second estimate = 25%), update was a negative value (M = 1.63, SD = 1.69).

Additionally, in order to obtain inter-individual measures of the optimism bias, we computed the difference between the mean update for positive and negative trials, separately

for self and other, for each participant. The self-specific optimism bias ($bias_{self}$) was computed as mean $update_{self\_p}$ – mean $update_{self\_n}$, and the other-related optimism bias ($bias_{other}$) was computed as mean $update_{other\_p}$ – mean $update_{other\_n}$. Positive values indicate an optimism bias as they result from larger mean updates after desirable than undesirable new information. Finally, a differential optimism bias ($bias_{self\_other}$) was computed as $bias_{self}$ – $bias_{other}$. Here, a positive value indicates a greater self-related than other-related optimism bias. In addition, we computed the difference between the mean first estimates ($1^{st}E$) relating to self and other ($1^{st}E_{self\_other}$): mean $1^{st}E_{self}$ – mean $1^{st}E_{other}$. The difference in initial risk estimates for adverse future events represents a classical measure for the 'comparative optimism' beyond update paradigms (Shepperd et al., 2002). These bias measures, $bias_{self}$, $bias_{other}$, $bias_{self\_other}$, and $1^{st}E_{self\_other}$ were used to test for correlations with self-report questionnaire scores.

2.5 Statistical Analyses

2.5.1 Behavioral data. All behavioral analyses were conducted by using IBM SPSS Statistics (version 2.0). In order to test for an optimism bias in belief updating under consideration of participants' trait optimism, we conducted a mixed ANOVA with two within-subject factors target (self vs. other) and valence (positive vs. negative), one between-subject factor trait optimism (low_to vs. high_to), and the dependent variable mean update (general linear model, repeated-measures design). In addition, in order to better characterize the effect of trait optimism, two repeated-measures ANOVAs were conducted with target and valence as within-subject factors, separately for the two groups of participants with low and high trait optimism, respectively. Pearson's correlation coefficient, *r*, was used as an effect size measure (all *F*-values had 1 degree of freedom).

In order to test for relationships between the individual measures of the optimism bias ($bias_{self}$, $bias_{other}$, $bias_{self\_other}$, $1^{st}E_{self\_other}$), trait optimism (LOT-R) and depression symptoms (BDI), Pearson's correlation coefficients (*r*) were calculated. We expected that $bias_{self}$, $bias_{self\_other}$ and $1^{st}E_{self\_other}$ would relate positively to each other and to LOT-R, and negatively

to BDI corresponding to reduced optimism in depression (Korn et al., 2014; Scheier et al., 1994; Strunk and Adler, 2009; Strunk et al., 2006), thus one-tailed tests were applied. We did not have any hypotheses about the relationship between $bias_{other}$ and these scores, so we applied two-tailed tests here. The alpha level of .05 was used for all statistical tests.

2.5.2 FMRI data. FMRI data were preprocessed and analyzed using MATLAB 7.9 (The MathWorks, Inc., Natick, MA, USA) and Statistical Parametric Mapping software package (SPM8, Wellcome Trust Center for Neuroimaging, London, UK). The EPI images were corrected for head movements using realignment and unwarping, normalized to the Montreal Neurological Institute reference space using the unified segmentation function, resampled to 2 mm × 2 mm × 2 mm voxels, and spatially smoothed with an 8 mm full width at half maximum Gaussian kernel.

Two separate analyses were conducted: i) a linear parametric modulation of the neural activity during the *2nd estimation* by update, and, ii) a linear parametric modulation of the neural activity during the *presentation of base rates* by update (see S.6 for an illustration and more details). The first analysis aimed to identify brain regions where the activity during *belief updating* correlated with the size of the update, and the second analysis explored whether there were brain regions, in which the activity during the *confrontation with disconfirming new information* correlated with – and thus predicted – subsequent update sizes.

At the single-subject level, conditions were modeled using a boxcar reference vector convolved with the canonical hemodynamic response function and its time derivative. Events relating to first estimation (duration 3 s, including target instruction), base rates (duration 2 s) and second estimation (duration 2 s) were separately modeled for the different experimental conditions (see Figure 1 and Figure S.1). Button presses indicating the first and the second estimate were modeled on separate regressors (duration from the onset of the response event to the last button press). Trials with missing responses, and trials with estimation error of

zero, if present, were modeled on a separate regressor. Movement parameters were included as multiple regressors of no interest. Low-frequency signal drifts were filtered using a cutoff of 128 s. Fourteen contrast images were computed relative to the implicit baseline (i.e., weighting a single regressor of interest with 1 in both sessions and the rest of regressors with 0) and used for group-level models: 1st estimation: self and other; base rate: self_p, self_n, other_p, other_n; 2nd estimation: self_p, self_p_PM (PM, parametric modulation by update) self_n, self_n_PM, other_p, other_p_PM, other_n, other_n_PM. In the second analysis applying PM to the average BOLD response to base rates, PM-contrasts referred to base rate presentations and not to 2nd estimations (see Figure S.1).

At the group level, for each of two abovementioned analyses, a flexible factorial design with factors 'condition' and 'subject' was conducted. The threshold for significance was set to $p < .05$, familywise error (FWE) -corrected at the voxel level for the whole brain with an extent threshold of 10 voxels. Because the main interest concerned neural correlates of varying update sizes, we report all significant results relating to whole brain parametric modulation effects from both analyses. In order to explore whether there are differences in the BOLD response pattern between the low_to and the high_to participants, beta values were extracted for all significant cluster maxima and compared between the subsamples using a t-test ($p < .05$, not corrected for multiple comparisons). For completeness, we also report whole brain main effects relating to average BOLD responses, i.e., independent of parametric modulation by update (see S.7). Furthermore, we also report and briefly discuss behavioral and neural results relating to estimation errors (see S.4, S.8, S.10). Activations were displayed on sections of the mean normalized T1 image of the overall sample. Anatomical labels were assigned using the Anatomy toolbox (Amunts et al., 2007; Eickhoff et al., 2005).

3 Results

3.1 Behavioral Data

3.1.1 Differential updating dependent on target, valence and trait optimism. The behavioral results of the mixed ANOVA show that there was a significant interaction between target and valence: Participants updated less in response to undesirable than to desirable base rates, and this effect was stronger for self-related than for other-related judgments (see Table 1 and Figure 2A). However, there was a significant three-way interaction indicating that the groups with low and high trait optimism demonstrated different update behavior patterns. The additional two ANOVAs (separately conducted for the low_to and high_to subsamples) revealed that the target × valence interaction was significant only in the high_to group, but not in the low_to group (see Table 1 and Figure 2B). Thus, only the group with high trait optimism showed a significant self-specific optimism bias. The pattern of the update behavior in the high_to group can also be described as a *selectively decreased update behavior during self_n* [self_n vs. self_p see Table 1; *p*s < .01 for self_n vs. other_p and self_n vs. other_n], relative to all other conditions with a comparably high update level (all *p*s > .05 for self_p vs. other_p, self_p vs. other_n, and other_p vs. other_n).

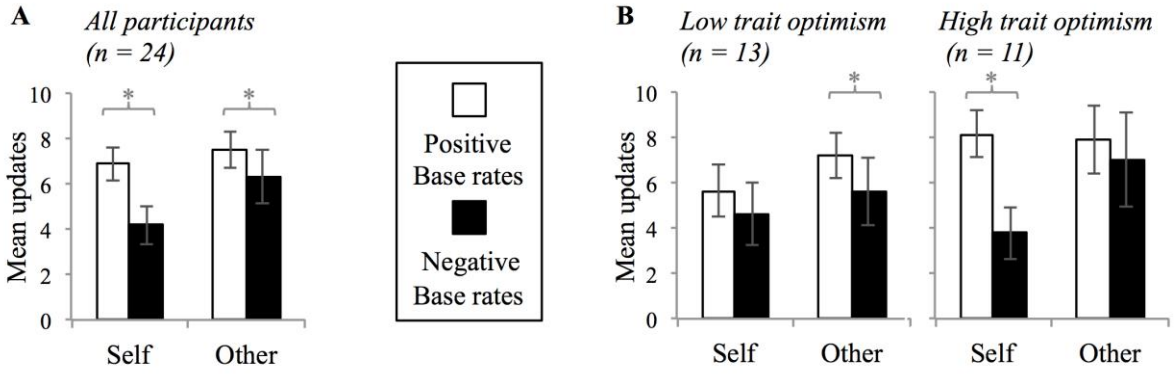

Figure 2 Behavioral results of the belief update experiment.
A) The overall sample shows the expected interaction between target person and valence of base rates: Participants generally tended to update less after undesirable (negative) new information, and this tendency was greater for self-related than for other-related judgments indicating an optimism bias.
B) High and low scorings on trait optimism were associated with a differential pattern of the update behavior. The self-protective optimism bias was strongly pronounced in participants scoring high on trait optimism (right), but absent in participants scoring low on trait optimism (left). Error bars show 95% C.I. * $p < .05$.

The reported ANOVA results could be replicated when using update values standardized to a mean of 0 and a standard deviation of 1 within each subject taking into

account inter-individual differences in the mean tendency to update initial estimates (group × target × valence interaction: $p = .001$; see S.4 for more details). In addition, confirming the differential updating pattern between subsamples scoring high and low on trait optimism, an ANCOVA based on the overall sample and including target person and valence as within-subject factors and LOT-R scores as a covariate revealed a significant three-way interaction target × valence × LOT-R, $F(1,22) = 5.75$, $p = .025$.

Other task-related variables, including first and second estimates, base rates and estimation errors differed significantly across the four conditions [Fs(3,69) > 16, ps < .001], which was partly forced by the experimental design (for descriptive statistics and detailed explanations see S.3 and S.1). For instance, it is to be expected that base rates in the positive condition are lower than in the negative condition because they were generated by subtracting values from participants' first estimates, while negative base rates were computed by adding values to first estimates. Importantly, however, none of these variables had a significant relationship to update sizes (first estimate, $r = .12$, $p = .229$; second estimate, $r = -.01$, $p = .998$; base rate, $r = -.19$, $p = .066$; estimation error, $r = -.15$, $p = .159$; not corrected for multiple comparisons), and the ANOVA effects of target person, valence and trait optimism on update remained significant even after controlling for these variables (ps < .05; see S.5 for more details).

3.1.2 Relationships between bias measures and questionnaire scores. The individual bias values as well as the trait optimism and BDI scores are presented in Table 2. The correlation analysis revealed that both $bias_{self}$ and $bias_{self\_other}$ were positively related to trait optimism ($bias_{self}$: $r = .35$, $p = .046$; $bias_{self\_other:}$ $r = .46$, $p = .013$), and inversely related to depression symptoms ($bias_{self}$: $r = -.42$, $p = .020$; $bias_{self\_other:}$ $r = -.50$, $p = .007$). There was no significant correlation between $bias_{other}$ and trait optimism ($r = -.15$, $p = .475$) or depression symptoms ($r = .12$, $p = .574$). In addition, $1^{st}E_{self\_other}$ did not correlate with trait optimism ($r = .03$, $p = .447$), depression symptoms ($r = -.04$, $p = .436$), nor $bias_{other}$ ($r = .10$, $p = .656$).

However, there was a correlation between $1^{st}E_{self\_other}$ and $bias_{self}$ (r = -.36, p = .042), and $1^{st}E_{self\_other}$ and $bias_{self\_other}$ (r = -.42, p = .021). Furthermore, there was a strong inverse correlation between the trait optimism and depression symptoms (r = -.67, p = .000).

3.1.3 Post-experimental assessments. The analysis of the ratings of personal experience with stimulus events revealed that the stimulus events were equally familiar within the four experimental conditions [F (3,69) = 1.76, p = .163; overall: M = 3.00, SD = 0.36], and that participants indeed followed the instruction not to respond to events that they were experiencing at the time of the experiment. The examination of the short post-experiment debriefing revealed that the task was not too difficult, and that none of the participants who were included in the analyses were aware of the task purpose, namely, investigating the optimism bias as indicated by larger updates after desirable than after undesirable new information, in particular for self-related judgments. Participants reported that some base rates appeared surprisingly low or high, but that they attributed this to the fact that the probabilities related to the whole population and to the entire lifetime. None of the participants included in the analyses doubted the authenticity of the base rates. In fact, even after being informed about the task purpose and the manipulation of the base rates at the end of the experiment, none of the included participants declared that they were aware of either of these.

3.2 FMRI data

FMRI analyses aimed to identify brain regions in which the activity correlated with the size of updates on a trial-by-trial level, i.e., targeting within-subject effects. For the complete list of results relating to the parametric modulation by updates see Table 3.

3.2.1 Neural correlates of updates during second estimation. There was a significant difference between regression slopes for self_p and self_n in the vmPFC. The beta estimates in the Figure 3 show that the correlation was positive for self_p, but negative for self_n. That means that when making self-referential judgments, the activity in the vmPFC was greater the *larger* the *desirable* updates, and it was also greater the *smaller* the *undesirable* updates. This

effect also remained significant when conducting a paired t-test at the second level of analysis directly comparing the parametric regressors relating to self_p and self_n, respectively (T = 4.33, p_{uncorr} = .001, peak at 0/ 34/ -14, cluster size 36), however at a less stringent significance threshold. In addition, when repeating the same analysis, but with within-subject standardized update values instead of row update values, the self_p > self_n effect in the vmPFC was replicated (T = 5.51, $p_{FWE\text{-}corr}$ = .003, peak at -2/ 36/ -10, cluster size 50).

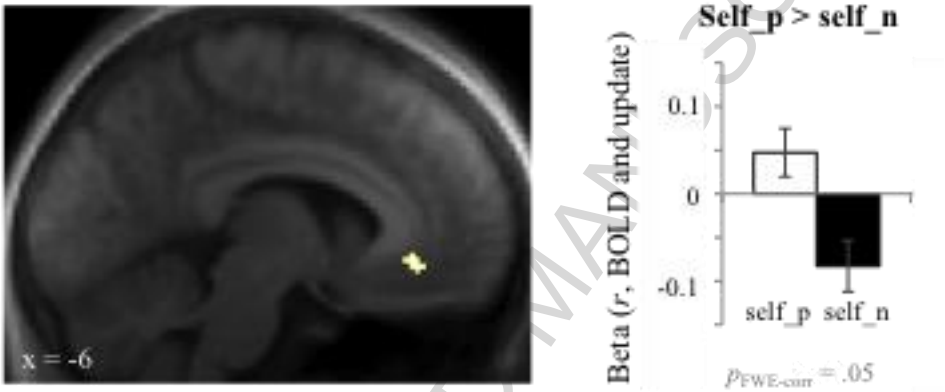

Figure 3 Activity in the subgenual ventromedial prefrontal cortex (vmPFC) during self-related belief updating reflected the subjective value of belief updates.
The direct comparison between self-trials with positive and negative base rates revealed a differential within-subject correlation between the activity in the vmPFC and the size of belief updates. After positive base rates (self_p) there was a positive correlation between the BOLD response and the size of updates: The larger the updates, the stronger the BOLD response. Conversely, after negative base rates (self_n) there was an inverse correlation between the BOLD response and the size of updates: The smaller the update, the stronger the BOLD response. This activation pattern reflects the subjective value of updating: Both large updates leading to an improvement of future outlooks relative to the initial belief, and small updates (or refraining from updating) towards worse future outlooks are expected to have a positive subjective value, and were accompanied by an increased BOLD response in the vmPFC. Error bars show 90% C.I.

No comparable effect was present for other-related judgments (other_p > other_n). Moreover, the interaction effect [(self_p > self_n) > (other_p > other_n)] yielded a large activation in the vmPFC (maximum peak at 14/ 42/ -4, T = 5.17, size = 1621 voxels), including a peak at -6/ 34/ -6 (T = 3.95), however at a more liberal threshold ($p_{FWE\text{-}corr}$s < .05 at the cluster level, with p < .001 at the voxel level). This interaction confirms that the difference in regression slopes for positive and negative base rates was indeed specific for self-related judgments.

3.2.2 Neural correlates of updates during base rate presentation. In *self*-trials with *undesirable* base rates (self_n), there was an inverse correlation between the BOLD response to base rates and updates in a network including the thalamus, the fusiform gyrus extending into the hippocampus, occipital areas, vStr and the dmPFC (see Table 3B and Figure 4). Thus, during the reception of undesirable self-related base rates, the activity in these regions was greater the lower the size of subsequent updates.

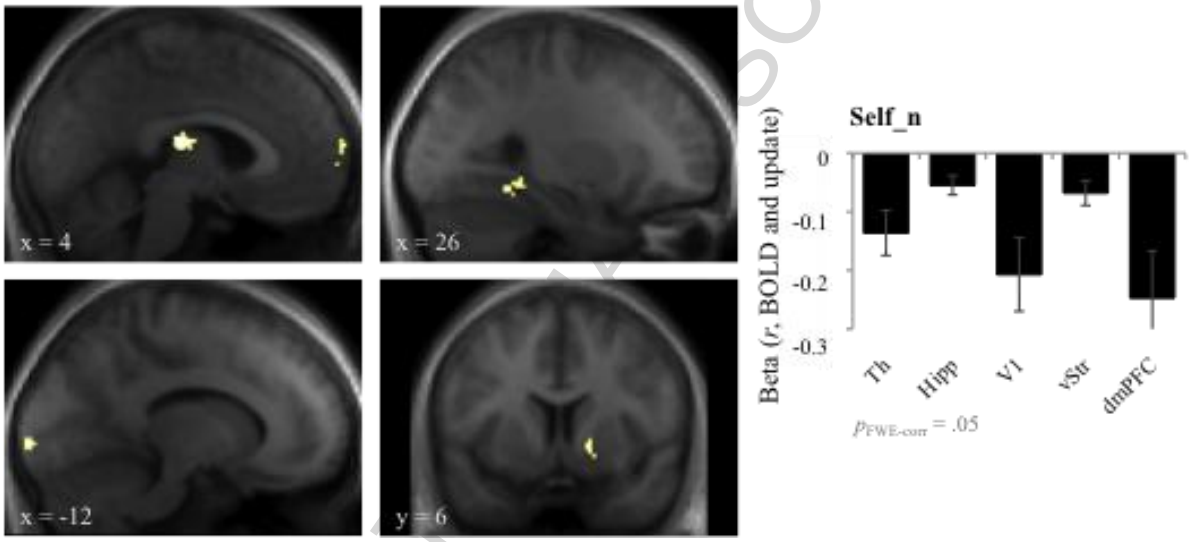

Figure 4 Neural activity during the presentation of self-related base rates that correlated with the size of subsequent updates on the trial-by-trial level.
There was an inverse correlation between the BOLD response to base rates and the subsequent updates only after self-related negative base rates (self_n). The greater recruitment of the network including the thalamus (Th), the fusiform gyrus extending into the hippocampus (Hipp), the primary visual cortex (V1), the ventral striatum (vStr) and the dorsomedial prefrontal cortex (dmPFC), the smaller the subsequent belief updates resulting in an undesirable future outlook.

The reported self_n effect also remained significant in the majority of identified regions when conducting a one sample t-test at the second level of analysis including the parametric regressor relating to self_n (dmPFC, thalamus, lingual gyrus, vStr; p_{uncorr}s > .002, *T*s > 3.3), however at a less stringent significance threshold. In addition, when repeating the same analysis, but with within-subject standardized update values instead of raw update values, we could replicate the inverse correlation effect for self_n in all regions listed in Table 3B except for the dmPFC ($p_{FWE\text{-}corr}$s < .022).

No comparable effect was present for other-related judgments (other_n, inverse correlation). Moreover, the direct comparison (other_n > self_n) yielded the same network as the inverse self_n contrast alone, however at a more liberal threshold (p_{uncorr}s < .001, *T*s > 3.7). This comparison indicates that the increasing activation of the described network with decreasing subsequent updates during the processing of negative base rates was indeed specific for self-related judgments.

3.2.3. Control variables. In order to examine whether the update-related effects might be related to the size of presented base rates, or to the size of estimation errors, we computed supplementary analyses with parametric modulation by base rates and estimation errors, respectively, instead of updates (see Tables S.4 and S.5). Relative to the main results of parametric modulation by updates reported in 3.2.1 and 3.2.2, these analyses revealed no similarities or overlaps, indicating that the update-related effects cannot be accounted for by the influence of the base rates or estimation errors.

Moreover, because of the correlation between updates and estimation errors (see Table S.1), we examined more closely whether the differential BOLD signals reported in 3.2.1 and 3.2.2 were better accounted for by updates than by estimation errors. For each phase of the trial (second estimation and base rate presentation), we computed two separate first level models (Wood et al., 2008). One model included updates as parametric modulators, and the other model included estimation errors as parametric modulators. In order to scale the estimated regression parameters uniformly, the parameters representing estimation errors and updates were standardized to a mean of 0 and a standard deviation of 1 (Wood et al., 2008). Note that we did not include updates and estimation errors simultaneously into one model because of high correlations between these two variables. The resulting multicollinearity would have compromised the estimation of model parameters and the interpretability of separate parametric modulators (Wood et al., 2008).

At the second level of the comparative analysis relating to *second estimation*, a flexible factorial model with two regressors relating to parametric modulation of self_p and self_n by updates, and with two regressors relating to parametric modulation of self_p and self_n by estimation errors were computed, and the vmPFC cluster from Table 3A was used as search volume. In accordance with the finding that only update but not estimation error was a significant predictor of the differential activity in the vmPFC, the interaction effect [(self_p_{update} > self_n_{update}) > (self_$p_{estimation\ error}$ > self_$n_{estimation\ error}$)] revealed that the effect of updates on differential activity in the vmPFC was significantly stronger than the one of estimation errors (T = 2.77, $p_{FWE\text{-}corr}$ = .030, peak at -6/ 38/ -8, cluster size 34).

At the second level of the comparative analysis relating to *base rate presentation*, a flexible factorial model with two regressors relating to parametric modulation of self_n by updates and estimation errors, respectively, was computed and the significant clusters from Table 3B (inverse contrast self_n) were used as search volume. In accordance with the finding that only update but not estimation error was a significant predictor of the activity in the respective areas, the comparison self_n_{update} < self_$n_{estimation\ error}$ revealed that the effect of updates was significantly stronger than the one of estimation errors in the lingual gyrus (T = 4.53, $p_{FWE\text{-}corr}$ = .002, peak at -22/ -68/ -6, cluster size 22), and the vStr (T = 4.25, $p_{FWE\text{-}corr}$ = .004, peak at 14/ 8/ -4, cluster size 13). For the thalamus, the parahippocampal gyrus, and the dmPFC the difference was significant only at a threshold not corrected for multiple comparisons (p_{uncorr}s < .05, Ts > 1.96).

Note that as we used the results reported in 3.2.1 and 3.2.2 as search volumes, the comparative analyses at the second level are biased in favor of analyses based on updates. Importantly, however, the comparative analyses were not conducted to examine whether updates or estimation errors can generally better account for the data, but to show that the update-dependent activation patterns are indeed significantly better accounted for by updates than by estimation errors.

3.2.4 Differential neural correlates of updates dependent on trait optimism.

Comparisons of the BOLD response patterns in all reported local maxima between subsamples with low and high trait optimism revealed only one significant difference, relating to the activity in the Rolandic operculum extending into the insula that exhibited an inverse correlation with updates (see Table 3, self_p, inverse correlation).

4 Discussion

The overall *behavioral* data corroborate prior findings of decreased updating in response to undesirable new information, and indicate that this effect is stronger for judgments referring to oneself than to others (Kuzmanovic et al., 2015; Sharot et al., 2011). The interaction effect reflecting the particularly strong optimism bias for self-referential judgments was also significant when using standardized update values, which control for inter-individual differences in the general tendency to update initial estimates. Moreover, the self-specific optimism bias in updates was independent of the size of first and second estimates, as well as of the size of presented base rates and induced estimation errors.

Debriefing revealed that participants were not aware of their biased belief updating. More specifically, none of the participants were aware that the difference in updating after desirable and undesirable information matters for the research question and that they generated such a difference in their own updating. This indicates that the optimism bias in belief updating represents a covert and unintentional influence on judgments. This spontaneous dismissal of negative information calling for unfavorable belief adjustments constitutes a possible mechanism for maintaining optimism in the face of challenging new evidence (Hughes and Zaki, 2015; Sharot et al., 2011), which is particularly important when beliefs are self-relevant.

Finally, behavioral results showed that the update behavior was modulated by trait optimism. Only highly trait-optimistic participants, but not those scoring low on trait optimism, showed the self-protective update pattern that selectively neglects *undesirable self-*

referential information, without at the same time exhibiting such a bias for other-related judgments. Along with the cross-correlations between trait optimism, self-reported depression symptoms, and biased updating, these results support the view that optimistic distortions in judgment formation are related to affective states and to general expectations towards the future. However, there was no significant correlation between trait optimism and the initial tendency to estimate one's own risks for adverse events to be lower than those of similar others (i.e., before being provided with base rates, $1^{st}E_{self\_other}$), in good accordance with the literature referring to 'comparative optimism' (Shepperd et al., 2002).

The primary aim of the *fMRI data analyses* was to provide empirical support for the assumed motivational causes of the optimism bias. Such motivational explanations relate to the assumed pleasure of favorable self-referential prospective thinking, and to decision processes guided by these desired end-states of judgments (D'Argembeau et al., 2009; Hughes and Zaki, 2015; Sharot et al., 2011; Shepperd et al., 2002). In contrast to previous research launching the present paradigm (Sharot et al., 2011), we focused on the neural correlates of *updates*, and not on those of *estimation errors*, because updates are expected to reflect the *subjective value* of the adjusted expectation towards the future more closely than estimation errors. For completeness, we report and discuss replications of previous results relating to estimation errors (Sharot et al., 2011) in the Supplementary Material (S.4, S.8 and S.10). Importantly, we could show that the effects in the vmPFC and the vStr represent unique update-dependent variance in the BOLD signal, which is different from the one relating to estimation errors, or base rates.

In the following, we discuss the neural correlates of updates during the second estimation and during the reception of base rates, respectively. While it is not possible to precisely determine the point of updating during the period from the presentation of the base rates up to the response indicating the final estimate, we roughly assume that the presentation of base rates represents an early and preliminary stage of the judgment formation process,

while the phase immediately before the response indicating the second estimate represents the final decision.

4.1 Neural correlates of updates during the second estimation

The differential subjective value of *self-referential* belief updates at the time point of *the final update decision* was reflected by the activity in the subgenual vmPFC: The BOLD response increased both with increasing favorable and with decreasing unfavorable updates. In other words, both large favorable belief updates, and small unfavorable belief updates (or refraining from updating) were associated with increased activity in the vmPFC. Critically, this activation pattern was not present during judgments regarding similar others, and the interaction term confirmed that it was specific for judgments concerning oneself (however at a more liberal threshold, see 3.2). Note that for both favorable and unfavorable updating, the size of updates represents the adjustment of the first estimate towards the presented base rate, which in both conditions results (mostly) in positive signed values (i.e., after positive/lower base rates, a lower second estimate was expected, thus update was computed as first estimate – second estimate, and vice versa). From the point of view of the judging subject, both large belief updates towards smaller risks, and small or zero belief updates towards higher risks have a positive value. Thus, the vmPFC reflected the particularly salient subjective value of belief updates relating to one's own future risks.

While the vmPFC has been shown to play a central role in representing the subjective value of emotional stimuli and rewards (Levy and Glimcher, 2012; Ochsner et al., 2012; Winecoff et al., 2013), it also has been associated with other cognitive functions, for instance with encoding of personal significance (D'Argembeau, 2013; Kim and Johnson, 2015). Nevertheless, two reasons speak for the interpretation implying the representation of values. First, our findings are highly specific: the activity in the vmPFC increased with increasing favorable updates, and decreased with increasing unfavorable updates. Explanations relating to personal significance cannot account for these conversely directed results because

increasing sizes of both favorable and unfavorable updates should be of similar personal significance. Second, favorable belief updating recruited specifically the subgenual region of the vmPFC (also termed subgenual or subcallosal cingulate), which exactly overlaps with the only correlate of *expected value* reported by the most recent meta-analysis of human neuroimaging studies using reinforcement learning models (Chase et al., 2015, peak coordinates 4/ 34/ -6; see also Levy and Glimcher, 2012). In addition, probabilistic frameworks using large-scale mappings between human neural and cognitive states confirm that the subgenual vmPFC is selectively, and not just consistently associated with the term 'value' (http://www.neurosynth.org, z-score of the reverse inference map at -6/ 36/ -6 = 5.41; Yarkoni et al., 2011). Therefore, it can be argued that the differential engagement of the subgenual vmPFC supports the notion that favorable beliefs are associated with pleasure and emotional benefits, comparable to rewards.

Recently, the subgenual vmPFC has been related to motivational influences on *self-referential* cognition beyond the general representation of positive values and rewards. At the between-subject level, trait optimism scores correlated with the differential activity in the vmPFC while imagining self-referential positive future events relative to negative ones (Sharot et al., 2007). The vmPFC was also recruited by processing positive personality traits relative to negative ones (Beer and Hughes, 2010), particularly when they were regarded as self-descriptive relative to non-descriptive (Moran et al., 2006), or were ascribed to close vs. non-close others (Hughes and Beer, 2012). However, while these studies on self-enhancing processing were always compounded with the general effect of valence of stimuli (e.g., positive vs. negative traits or events), our study is the first to demonstrate that the subgenual vmPFC tracks the affective meaning of judgments, independently of the valence of judged stimuli. The subjective value in the present task is defined by trial-to-trial fluctuations of belief updating, which were always related to risks of experiencing negative future events. In such an overall negative context of reasoning, the subjective value of the final belief is

computed relative to the avoided, relatively worst alternative. In accordance with this, a computational modeling study recently demonstrated that successful avoidance of options associated with a high probability of punishment (75% probability of losing 0.5€) in favor of options associated with a lower risk of punishment (25% probability of loss) can acquire a positive value, although the chosen and the avoided options are both not rewarding per se (Palminteri et al., 2015). Similarly, in the context of stimuli with positive and negative valences, the vmPFC activity correlated with personal, self-reported ratings rather than with objective stimulus characteristics (Winecoff et al., 2013). Thus, the subgenual vmPFC seems to be involved in internal computation of values of beliefs, even in the absence of the actual satisfying (future) event outcome, and irrespective of the presence of rewarding stimuli per se (Leary, 2007). In turn, such internally modulated subjective values may play a critical role in emotion regulation and positively biased judgment formation (Ochsner et al., 2012).

Additional – yet indirect – support for the specific role of the subgenual vmPFC in motivated cognition during self-referential judgments comes from depression research. Particularly the subgenual vmPFC shows critical dysfunctions in patients with major depression, and is the target region of deep brain stimulation in treatment-resistant patients (Holtzheimer et al., 2012; Lozano et al., 2008). At the same time, depression has been associated with decreased optimism bias in belief updating (Garrett et al., 2014; Korn et al., 2014), and impaired emotion regulation (Joormann and Quinn, 2014). Thus, adaptive interactions between the mood state, belief formation, and reward and emotional reactivity seem to critically depend on the dense neural connections crossing the subgenual vmPFC.

At the time of *the final update decision*, additional brain regions in which the activity inversely correlated with the size of updates were revealed. Because these findings are not directly related to our research question, we do not discuss them in detail.

4.2 Neural correlates of updates during the reception of base rates

Our second analysis of the fMRI data focused on the correlation between updates and the BOLD signal *during the base rate presentation*. This analysis aimed to explore which neural activity predicts the size of the subsequent estimate adjustment at the moment of being confronted with the disproving information. For *self-related* judgments and *undesirable* base rates there was an inverse correlation between the subsequent update size and the activity in a network of regions including the thalamus, the fusiform gyrus extending into the hippocampus, early visual regions, the vStr and the dmPFC. In other words, the more participants recruited this network while being confronted with self-related undesirable base rates, the less they subsequently updated their initial estimate. A direct comparison confirmed that the effect was specific for judgments relating to oneself relative to those about others (however at a more liberal threshold, see 3.2). Consequently, this network may play a crucial role in dismissing the relevance of self-relevant unfavorable information in the early stage of the judgment formation process.

However, the exact cognitive processes involved in suppressing belief updates in response to undesirable information cannot be deducted from our data, as we did not assess specific thoughts or judgment criteria of the participants. Based on theoretical frameworks of comparative thinking (Mussweiler, 2003) and prior findings regarding the selection of the "average target" (Shepperd et al., 2002) we can make some suggestions regarding possible interpretations of the predictive coding of unfavorable updates. Possibly, personal characteristics (e.g., family history, life style and precautionary intentions) are selectively recalled in a way to support a relatively decreased personal risk compared to the average person. For instance, when confronted with an unexpectedly high base rate for a disease, one could recall one's own healthy life style that contrasts the overall unhealthy life style of the "average person" to promote the conclusion that high risks do not need to be applied to one self. This speculation is in good accordance with the known functional roles of the recruited network. The hippocampus (Wimmer and Shohamy, 2012) and the dmPFC (Spreng and

Grady, 2010; Spreng et al., 2009) are associated with mnestic, inferential and prospective processing necessary for case specific reasoning. Furthermore, the vStr and the thalamus play an important role in reward learning, which suggests that their involvement may reflect the adjustment of the information processing in line with the anticipated value of the ensuing belief updates (Galvan et al., 2005).

The predictive engagement of the vStr in the early stage of judgment formation and that of the vmPFC in its final phase fit well with the differential functions suggested for these core components of the neural reward circuitry. While the vStr has been related to prediction errors and to formation of preferences early after the confrontation with the target stimulus, the vmPFC has reflected the conceptual, abstract value of outcomes (Chase et al., 2015; Kim et al., 2007; O'Doherty, 2004; Roy et al., 2012). Thus, in order to dismiss information calling for unfavorable future outlooks, vStr may provide rapid signals that are able to interact with and influence other cognitive processes in order to achieve more valuable subsequent decision outcomes (Kim et al., 2007). In addition, the previous demonstration that such coding of initial preferences by the vStr occurs automatically, i.e., even in the absence of the explicit instruction to evaluate the value of a stimulus (Kim et al., 2007), is consistent with the fact that participants in our study were not aware of their selective neglect of undesirable, self-relevant base rates during their belief updating.

In contrast to self-relevant judgment formation, undesirable information relevant for judgments about *others* engaged a network adjacent to the white matter with significant but weak inverse correlations with subsequent updates.

5 Limitations

While we found significantly different behavioral patterns of belief updating between subsamples with high and low trait optimism, comparing these subsamples with respect to their neural correlates did not reveal significant differences (with one exception, contrast self_p, inverse correlation, see Table 3A). This indicates that the described neural responses

to updates varying in size and desirability represent a common feature of processing self-referential information, independent of inter-individual differences in general expectations towards the future. Another possible explanation is that non-significant results may be related to the inflated type II error due to the small subsample sizes recruited in our study. In addition to this general problem, the low variance of trait optimism may also explain the lack of significant differences in the neural correlates. However, recruiting participants with low trait optimism and negative self-views might require a systematic search because about 70% of the general population tend to have positive self-views (Swann and Jennifer, 2010).

Two other limitations of the present study relate to the estimation errors. First, our algorithm for manipulation of estimation errors could not achieve perfectly balanced errors between the conditions because mean positive estimation errors were lower than the negative ones (for more details see also S.1). While the range of the manipulated errors was still better controlled than if we had taken true and fixed base rates, the algorithm should be improved to enable fully balanced error computation. Critically, the interpretation of our behavioral results in terms of an optimistic bias is not questioned by this imbalance, because formal learning models predict that the updating should be greater the bigger the estimation error. Although in positive trials estimation errors were on average smaller, the mean belief updating was significantly larger. Consequently, the smaller estimation errors in positive trials indicate that the optimism bias may even be underestimated to some extent.

The second limitation relates to the imprecise assessment of subjectively experienced estimation errors (see also SI2 and Kuzmanovic et al., 2015). While formally the estimation error corresponds to the difference between participants' first estimate and the presented base rate in each trial of the experiment, the participants may not perceive this difference as an indication that their initial judgment was erroneous because of personal vulnerabilities or resources. For instance, a strong family history of cancer may suggest a higher personal risk of suffering from cancer relative to the population base rate, so that a presentation of a lower

population base rate need not be perceived as an error. In order to further improve the methodological precision of the paradigm, subjective estimation errors should be more carefully assessed.

6 Conclusion

Different ways of assessing the optimism bias exist, yet little is known about the underlying psychological processes and neural mechanisms. Our findings suggest that particularly when making judgments regarding one's own future and being confronted with undesirable information that calls for belief adjustments towards unfavorable future outlooks, a network including the vStr, the thalamus, the hippocampus, and the dmPFC is activated in order to prevent detrimental updating. More specifically, the more these regions associated with reward-based learning and complex cognitive processing (including memory retrieval, mental visual emulation and inferential reasoning) were activated, the smaller the subsequent unfavorable update. During the final computation of self-related belief updates, the positive value of these salient decisions was mirrored by the subgenual vmPFC: Both large belief updates towards better future outcomes, and refraining from updates that would have led to worse future outcomes have a positive subjective value and were accompanied by an increased activity. The fact that central nodes of the neural reward circuitry were recruited dependent on the valence of self-referential judgments provides empirical evidence that cognitive and motivational processes are closely intertwined in optimistic distortions in reasoning. Our findings support the conclusion that the human cognitive system is motivationally predisposed to dismiss undesirable information and attend to desirable information due to the emotional value of the so achieved optimistic judgment outcomes.

Acknowledgments

Anneli Jefferson acknowledges the support of the Costs and Benefits of Optimism project as part of the Hope and Optimism funding initiative.

References

Amunts, K., Schleicher, A., Zilles, K., 2007. Cytoarchitecture of the cerebral cortex--more than localization. Neuroimage 37, 1061-1065; discussion 1066-1068.
Bartra, O., McGuire, J.T., Kable, J.W., 2013. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76, 412-427.
Beer, J.S., Hughes, B.L., 2010. Neural systems of social comparison and the "above-average" effect. Neuroimage 49, 2671-2679.
Carver, C.S., Scheier, M.F., 2014. Dispositional optimism. Trends Cogn Sci 18, 293-299.
Chase, H.W., Kumar, P., Eickhoff, S.B., Dombrovski, A.Y., 2015. Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. Cogn Affect Behav Neurosci.
Chowdhury, R., Sharot, T., Wolfe, T., Duzel, E., Dolan, R.J., 2014. Optimistic update bias increases in older age. Psychol Med 44, 2003-2012.
Clithero, J.A., Rangel, A., 2014. Informatic parcellation of the network involved in the computation of subjective value. Soc Cogn Affect Neurosci 9, 1289-1302.
D'Argembeau, A., 2013. On the role of the ventromedial prefrontal cortex in self-processing: the valuation hypothesis. Front Hum Neurosci 7, 372.
D'Argembeau, A., Lardi, C., Van der Linden, M., 2012. Self-defining future projections: exploring the identity function of thinking about the future. Memory 20, 110-120.
D'Argembeau, A., Renaud, O., Van der Linden, M., 2009. Frequency, characteristics and functions of future-oriented thoughts in daily life. Applied Cognitive Psychology 25, 96-103.
Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25, 1325-1335.
Garrett, N., Sharot, T., Faulkner, P., Korn, C.W., Roiser, J.P., Dolan, R.J., 2014. Losing the rose tinted glasses: neural substrates of unbiased belief updating in depression. Front Hum Neurosci 8, 639.
Haber, S.N., Knutson, B., 2010. The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology 35, 4-26.
Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., Berke, J.D., 2016. Mesolimbic dopamine signals the value of work. Nat Neurosci 19, 117-126.
Holtzheimer, P.E., Kelley, M.E., Gross, R.E., Filkowski, M.M., Garlow, S.J., Barrocas, A., Wint, D., Craighead, M.C., Kozarsky, J., Chismar, R., Moreines, J.L., Mewes, K., Posse, P.R., Gutman, D.A., Mayberg, H.S., 2012. Subcallosal cingulate deep brain stimulation for treatment-resistant unipolar and bipolar depression. Arch Gen Psychiatry 69, 150-158.
Hughes, B.L., Beer, J.S., 2012. Orbitofrontal cortex and anterior cingulate cortex are modulated by motivated social cognition. Cereb Cortex 22, 1372-1381.
Hughes, B.L., Zaki, J., 2015. The neuroscience of motivated cognition. Trends Cogn Sci 19, 62-64.
Joormann, J., Quinn, M.E., 2014. Cognitive processes and emotion regulation in depression. Depress Anxiety 31, 308-315.
Kable, J.W., Glimcher, P.W., 2009. The neurobiology of decision: consensus and controversy. Neuron 63, 733-745.
Kim, H., Adolphs, R., O'Doherty, J.P., Shimojo, S., 2007. Temporal isolation of neural processes underlying face preference decisions. Proc Natl Acad Sci U S A 104, 18253-18258.

Kim, K., Johnson, M.K., 2015. Activity in ventromedial prefrontal cortex during self-related processing: positive subjective value or personal significance? Soc Cogn Affect Neurosci 10, 494-500.
Korn, C.W., Sharot, T., Walter, H., Heekeren, H.R., Dolan, R.J., 2014. Depression is related to an absence of optimistically biased belief updating about future life events. Psychological Medicine 44, 579-592.
Kuzmanovic, B., Jefferson, A., Vogeley, K., 2015. Self-specific optimism bias in belief updating is associated with high trait optimism. Journal of Behavioral Decision Making 28, 281–293.
Leary, M.R., 2007. Motivational and emotional aspects of the self. Annu Rev Psychol 58, 317-344.
Levy, D.J., Glimcher, P.W., 2012. The root of all value: a neural common currency for choice. Curr Opin Neurobiol 22, 1027-1038.
Lozano, A.M., Mayberg, H.S., Giacobbe, P., Hamani, C., Craddock, R.C., Kennedy, S.H., 2008. Subcallosal cingulate gyrus deep brain stimulation for treatment-resistant depression. Biol Psychiatry 64, 461-467.
Moran, J.M., Macrae, C.N., Heatherton, T.F., Wyland, C.L., Kelley, W.M., 2006. Neuroanatomical evidence for distinct cognitive and affective components of self. J Cogn Neurosci 18, 1586-1594.
O'Doherty, J.P., 2004. Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr Opin Neurobiol 14, 769-776.
Ochsner, K.N., Silvers, J.A., Buhle, J.T., 2012. Functional imaging studies of emotion regulation: a synthetic review and evolving model of the cognitive control of emotion. Ann N Y Acad Sci 1251, E1-24.
Palminteri, S., Khamassi, M., Joffily, M., Coricelli, G., 2015. Contextual modulation of value signals in reward and punishment learning. Nat Commun 6, 8096.
Peters, J., Buchel, C., 2010. Neural representations of subjective reward value. Behav Brain Res 213, 135-141.
Rose, J.P., Endo, Y., Windschitl, P.D., Suls, J., 2008. Cultural differences in unrealistic optimism and pessimism: the role of egocentrism and direct versus indirect comparison measures. Pers Soc Psychol Bull 34, 1236-1248.
Roy, M., Shohamy, D., Wager, T.D., 2012. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends Cogn Sci 16, 147-156.
Scheier, M.F., Carver, C.S., Bridges, M.W., 1994. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. J. Pers. Soc. Psychol. 67, 1063-1078.
Sharot, T., Kanai, R., Marston, D., Korn, C.W., Rees, G., Dolan, R.J., 2012. Selectively altering belief formation in the human brain. Proc Natl Acad Sci U S A 109, 17058-17062.
Sharot, T., Korn, C.W., Dolan, R.J., 2011. How unrealistic optimism is maintained in the face of reality. Nat Neurosci 14, 1475-1479.
Sharot, T., Riccardi, A.M., Raio, C.M., Phelps, E.A., 2007. Neural mechanisms mediating optimism bias. Nature 450, 102-105.
Shepperd, J.A., Carroll, P., Grace, J., Terry, M., 2002. Exploring the causes of comparative optimism. Psychologica Belgica 42, 65-98.
Shepperd, J.A., Klein, W.M.P., Waters, E.A., Weinstein, N.D., 2013. Taking stock of unrealistic optimism. Perspectives on Psychological Science 8, 395-411.
Spreng, R.N., Grady, C.L., 2010. Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. J Cogn Neurosci 22, 1112-1123.

Spreng, R.N., Mar, R.A., Kim, A.S., 2009. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J Cogn Neurosci 21, 489-510.
Strunk, D.R., Adler, A.D., 2009. Cognitive biases in three prediction tasks: a test of the cognitive model of depression. Behav Res Ther 47, 34-40.
Strunk, D.R., Lopez, H., DeRubeis, R.J., 2006. Depressive symptoms are associated with unrealistic negative predictions of future life events. Behav Res Ther 44, 861-882.
Weinstein, N.D., 1980. Unrealistic Optimism About Future Life Events. Journal of Personality and Social Psychology 39, 806-820.
Weinstein, N.D., 1987. Unrealistic Optimism About Susceptibility to Health-Problems - Conclusions from a Community-Wide Sample. Journal of Behavioral Medicine 10, 481-500.
Weinstein, N.D., Klein, W.M., 1996. Unrealistic optimism: Present and future. Journal of Social and Clinical Psychology 15, 1-8.
Winecoff, A., Clithero, J.A., Carter, R.M., Bergman, S.R., Wang, L., Huettel, S.A., 2013. Ventromedial prefrontal cortex encodes emotional value. J Neurosci 33, 11032-11039.
Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 8, 665-670.

Tables

Table 1

Effects of target person and valence of new information on update behavior in the overall sample and in the sub-samples with low and high trait optimism

Source	*M*	*SD*	*F*	*p*	*r*
		All participants (n = 24)			
Target: self/other	5.38/6.90	1.49/2.02	20.42	.000	.69
Valence: p/n	7.14/5.27	1.70/2.19	17.50	.000	.67
Group: low_to/high_to	5.79/6.67	1.53/1.66	2.19	.153	.30
Target × valence			7.98	.010	
Group × target			0.12	.737	
Group × valence			2.00	.171	
Group × target × valence			17.25	.001	
Pairwise comparisons					
Self_p/self_n	6.79/4.24	2.14/2.03	24.62	.000	.73
Other_p/other_n	7.50/6.26	1.92/2.80	5.27	.032	.44
		Low trait optimism (n = 13)			
Target: self/other	5.09/6.43	1.56/1.78	12.39	.004	.71
Valence: p/n	6.50/5.11	1.57/2.19	4.41	.058	.52
Target × valence			0.91	.360	
Pairwise comparisons					
Self_p/self_n	5.63/4.63	1.92/2.27	1.68	.219	.35
Other_p/other_n	7.19/5.62	1.64/2.47	7.54	.018	.62
		High trait optimism (n = 11)			
Target: self/other	5.73/7.45	1.39/2.23	8.53	.015	.66
Valence: p/n	7.91/5.45	1.59/2.29	13.63	.004	.74
Target × valence			15.40	.003	
Pairwise comparisons					
Self_p/self_n	8.15/3.77	1.51/1.69	35.13	.000	.87
Other_p/other_n	7.87/7.02	2.22/3.08	0.84	.382	.27

Note: In order, *df*s for the three reported samples were 1, 22; 1, 12 and 1, 10. The overall sample was divided into a low and a high trait optimism sub-sample based on the median Life Orientation Test score (17). P and n, positive and negative valence of base rates, respectively.

Table 2

Descriptive statistics of optimism bias measures and questionnaire scores in the overall sample and in the sub-samples with low and high trait optimism

	All participants (n = 24)		*Low trait optimism (n = 13)*		*High trait optimism (n = 11)*	
	Min/Max	*M (SD)*	*Min/Max*	*M (SD)*	*Min/Max*	*M (SD)*
$Bias_{self}$*	-4.54/8.00	2.55 (3.11)	-4,54/5.03	0.70 (2.69)	-0.31/8.00	4.20 (0.79)
$Bias_{other}$	-3.87/4.47	1.24 (2.54)	-1.86/3.81	1.47 (2.12)	-3.87/4.47	0.51 (3.01)
$Bias_{self\_other}$*	-2.91/8.54	1.31 (3.26)	-2.91/4.44	-0.77 (2.11)	-1,52/8.54	3.69 (3.09)
$1stE_{self\_other}$	-19.67/2.45	-7.29 (5.86)	-19.67/2.45	7.29 (6.53)	-17.23/0.37	7.29 (5.27)
LOT-R*	8/22	17.00 (3.60)	8/17	14.33 (2.90)	18/22	20.10 (1.45)
BDI*	0/11	3.88 (3.14)	1/11	5.58 (3.09)	0/5	1.70 (0.65)

$Bias_{self}$ = mean updates after desirable information – mean updates after undesirable information, in self-trials; $Bias_{other}$ = mean updates after desirable information – mean updates after undesirable information, in other-trials; $Bias_{self\_other} = bias_{self} – bias_{other}$; $1stE_{self\_other}$, 1^{st} $estimate_{self} – 1^{st}$ $estimate_{other}$; LOT-R, Life Orientation Test, assessing trait optimism on a scale from 0 (pessimistic) to 24 (optimistic); BDI, Beck Depression Inventory, assessing symptoms of depression on a scale from 0 (minimal) to 63 (maximal). The overall sample was divided into a low and a high trait optimism sub-sample based on the median LOT-R score. * $p < .01$ for *t*-tests testing for differences between the sub-samples with low and high trait optimism.

Table 3

Brain regions in which the activity correlated with the size of updates: A) during the updating and B) during the base rate presentation

Cluster level	Voxel level							Subsample comparisons			
Size	Anatomical label †		*p* FWE-corr	*t*	x	y	z	ß *low_to*	ß *high_to*	*t*	*p*
A) Activity during 2nd estimation correlating with the size of updates											
	self_p > self_n										
38	vmPFC	M	.001	5.73	-6	34	-6	0.11	0.09	0.34	.739
	self_p, inverse correlation										
3424	Calcarine gyrus (V1/V2)	L	.000	7.24	-4	-76	6	-0.21	-0.15	-0.85	.406
	Calcarine gyrus	R	.000	6.87	8	-78	4	-0.21	-0.09	-1.68	.107
563	TPJ	L	.000	6.63	-56	-40	28	-0.20	-0.10	-1.30	.206
294	R. operculum	L	.000	6.19	-44	-12	18	-0.13	-0.04	-2.52	.019
	Insula	L	.000	6.05	-36	-2	10	-0.13	-0.03	-1.76	.093
23	STS	L	.006	5.33	-52	-54	8	-0.09	-0.07	-0.27	.791
17	IPL (Area 2)	L	.014	5.13	-48	-26	42	-0.11	-0.08	-0.43	.675
	self_n, inverse correlation										
5240	Calcarine gyrus (V1/V2)	R	.000	7.93	14	-84	10	-0.28	-0.18	-0.96	.348
	Calcarine gyrus	L	.000	7.38	-16	-70	8	-0.19	-0.18	-0.12	.909
16	Middle temporal gyrus	R	.003	5.45	52	-34	-6	-0.16	-0.09	-1.26	.221
13	Middle occipital gyrus	L	.006	5.34	-40	-80	12	-0.12	-0.05	-0.77	.448
10	Middle temporal gyrus	L	.021	5.03	-60	-58	18	-0.21	-0.16	-0.78	.447
	other_p, inverse correlation										
3713	Calcarine gyrus (V1/V2)	L	.000	7.35	-6	-74	6	-0.19	-0.18	-0.24	.813
	Calcarine gyrus	R	.000	6.96	8	-82	10	-0.21	-0.23	0.24	.812
67	Supramarginal gyrus	L	.001	5.78	-64	-22	26	-0.15	-0.10	-0.64	.526
83	STS	L	.001	5.64	-58	-52	10	-0.10	-0.18	1.01	.324
20	Insula	R	.002	5.60	36	6	10	-0.12	-0.08	-0.70	.491
90	TPJ	L	.003	5.48	-54	-38	32	-0.10	-0.11	0.23	.820
24	Superior temporal gyrus	R	.006	5.33	60	-38	18	-0.11	-0.15	0.66	.515
	other_n, inverse correlation										
2225	SOG (V1/V2)	R	.000	6.60	16	-92	18	-0.16	-0.26	1.12	.273
	Calcarine gyrus	R	.000	6.42	8	-70	18	-0.09	-0.19	1.07	.296
	Calcarine gyrus	L	.000	6.37	-8	-78	22	-0.08	-0.19	1.54	.137
29	Lingual gyrus	R	.001	5.69	26	-46	-8	-0.07	-0.07	0.12	.904
26	TPJ	L	.003	5.51	-52	-38	16	-0.08	-0.07	-0.20	.846
24	Fusiform gyrus	R	.003	5.45	28	-62	-10	-0.05	-0.09	0.96	.347

B) Activity during base rate presentation correlating with the size of updates											
	self_n, inverse correlation										
72	Thalamus	M	.000	5.87	4	-18	14	-0.15	-0.19	0.53	.604
18	Fusiform gyrus	R	.001	5.64	28	-46	-12	-0.07	-0.08	0.24	.814
	Parahipp. gyrus	R	.002	5.55	26	-38	-10	-0.05	-0.05	-0.08	.940
24	Calcarine gyrus (V1)	L	.003	5.45	-12	-98	-2	-0.17	-0.26	0.57	.575
22	Lingual gyrus (V4)	L	.004	5.40	-24	-66	-6	-0.11	-0.09	-0.25	.807
13	vStr	R	.005	5.37	14	6	-4	-0.06	-0.12	1.04	.310
15	dmPFC	M	.019	5.07	4	66	10	-0.15	-0.30	1.01	.322
	other_p, inverse correlation										
119	Middle frontal gyrus	R	,000	6.41	22	12	34	-0.03	-0.04	0.92	.370
14	Inferior temporal gyrus	L	,001	5.63	-40	-22	-12	-0.04	-0.04	0.04	.968
16	Insula	L	,003	5.48	-30	-4	12	-0.05	-0.04	-0.42	.679
	other_n										
52	Middle frontal gyrus	L	,004	5.42	-26	6	52	-0.01	0.14	-1.55	.135

† Anatomical labels and assignments to functional areas (only reported when relating to more than 20% of the cluster) refer to the Anatomy toolbox. Area 2, primary somatosensory cortex; IPL, inferior parietal lobule; Parahipp., parahippocampal; R. operculum, Rolandic operculum; SOG, superior occipital gyrus; STS, superior temporal sulcus; TPJ, temporoparietal junction; L, left, R, right, M, medial. Coordinates refer to the Montreal Neurological Institute space. Subsample comparisons were based on mean beta values (ß) extracted from listed local maxima; low_to, subsample scoring low on trait optimism, n = 13; high_to, subsample scoring high on trait optimism, n = 11; self_p, self-related judgments and positive base rates; self_n, self-related judgments and negative base rates; other_p, other-related judgments and positive base rates; other_n, other-related judgments and negative base rates. There were no significant results for the following contrasts: 1) Activity during 2nd estimation correlating with the size of updates: positive correlations for self_p, self_n, other_p, other_n, self_n > self_p, other_p > other_n, other_n > other_p, interaction effects; 2) Activity during base rate presentation correlating with the size of updates: positive correlations for self_p, self_n, other_p, negative correlations for self_p, other_n, self_p > self_n, self_n > self_p, other_p > other_n, other_n > other_p, interaction effects.

Highlights

- Updates of beliefs about one's own future are optimistically biased
- Positive values of updates are tracked by the ventromedial prefrontal cortex
- Activity in the ventral striatum predicts the neglect of undesirable information
- Neural reward circuitry is involved in optimistically biased judgment formation
- Results support motivational explanations of the self-related optimism bias