Cognitive regulation of ventromedial prefrontal activity evokes lasting change in the perceived self-relevance of persuasive messaging

Doré, B.P, Cooper, N., Scholz, C., O'Donnell, M.B., Falk, E.B.

Annenberg School for Communication, University of Pennsylvania

Corresponding authors:
Bruce Dore (brucedore@gmail.com)
Emily Falk (falk@asc.upenn.edu)

Persuasive messages can change people's thoughts, feelings, and actions, but these effects depend on how people think about and appraise the meaning of these messages. Drawing from research on the cognitive control of emotion, we used neuroimaging to investigate neural mechanisms underlying cognitive regulation of the affective and persuasive impact of advertisements communicating the risks of binge drinking, a significant public health problem. Using cognitive control to up-regulate (versus down-regulate) responses to the ads increased: negative affect related to consequences of excessive drinking, perceived ad effectiveness, and ratings of ad self-relevance made after a one-hour delay. Neurally, these effects of cognitive control were mediated by goal-congruent modulation of ventromedial prefrontal cortex (vmPFC) and distributed brain patterns associated with negative emotion and subjective valuation. These findings suggest that people can leverage cognitive control resources to deliberately shape responses to persuasive appeals, and identify mechanisms of emotional reactivity and integrative valuation that underlie this ability. Specifically, brain valuation pattern expression mediated the effect of cognitive goals on perceived message self-relevance, suggesting a role for the brain's valuation system in shaping responses to persuasive appeals in a manner that persists over time.

Persuasive messages can succeed or fail depending not only on objective qualities of the message, but on the goals and mindset of the receiver. However, the brain mechanisms that underlie this phenomenon are not well understood. Here, we propose that people exposed to a persuasive message can deliberately shape their emotion- and value-related brain responses, and these altered brain responses in turn shape emotion and evaluations.

Social psychological theorizing suggests that, far from being passive consumers of persuasive content, people actively interpret information in their social environment through a process of subjective construal (Bruner, 1957; Griffin & Ross, 1991; Fujita et al., 2008). Accordingly, the success or failure of an attempt at persuasion can hinge not only on the content of the appeal itself but on the way this content is appraised and evaluated by its recipient. Following this insight, models of persuasion have long recognized a need to account for both bottom-up (message-driven) and top-down (goal-driven) influences on the effects of persuasive messages (Carey, 1989; Chaiken, 1980; Hovland, Janis, & Kelly, 1953; Petty & Cacioppo, 1986; Witte, 1992).

Responses to persuasive appeals emerge from multiple processes that unfold in parallel, including reactivity to the persuasive content, appraisal of its meaning, and perceptions of its relevance and value. Although the outcomes of these processes can be indexed after the fact with behavioral and self-report methods, neuroimaging measures are particularly well-suited to capturing their underlying mechanisms in the moment that persuasive effects take hold (Falk & Scholz, 2018). Together, emerging models of the brain systems underlying cognitive control, emotional reactivity, and integrative valuation can provide an organizing framework for understanding the

mechanisms by which people can shape their responses to persuasive messages. Following from theories of cognitive control that posit prefrontally-mediated flexibility in the mapping between environmental stimuli and attention, cognition, and behavior (Miller & Cohen, 2001; Petersen & Posner, 2012), a large body of research shows that cognitively regulating emotional responses engages a network of prefrontal and parietal brain regions associated with domain-general cognitive control, and can evoke goal-congruent changes in emotional responding (Buhle et al., 2014; Ochsner et al., 2012). In tandem, other work has shown the ventromedial prefrontal cortex (vmPFC) to be involved in integrating information from diverse subcortical and cortical brain regions into a summary signal of the subjective value of a stimulus (Bartra, McGuire, & Kable, 2013; Clithero & Rangel, 2014). Further, prior studies suggest a prominent role for vmPFC in persuasion, in that vmPFC responses to persuasive messages has been associated with subsequent message-consistent behavior change (Falk & Scholz, 2018). However, this work has not made clear whether vmPFC responses reflect stable message-level or person-level characteristics, or whether they are flexibly constructed in a goal-driven manner. In the present research, in line with theories of cognition-emotion interaction more generally (e.g., Pessoa, 2008) we integrate literatures on the cognitive control of emotion with the neural bases of persuasion to test an account of how people can cognitively shape the way they respond to persuasive appeals and identify the neural mechanisms supporting this capacity. Doing so is a critical step toward better understanding how people can deliberately shape the meaning they derive from persuasive appeals in a controlled and reflective manner.

Here, we propose that deliberately controlling one's response to a persuasive message relies on recruitment of a domain-general cognitive control system and evokes goal-congruent changes in systems associated with emotional reactivity and integrative valuation, which in turn shape subjective evaluations. To test this, we conducted a neuroimaging study in which we instructed participants to cognitively up- or down-regulate the way they appraised persuasive messages that communicate the risks of excessive alcohol consumption, a significant public health problem (NIAAA, 2015). We examined message-related responses in vmPFC as well as multivariate patterns associated with valuation (Bartra, McGuire, & Kable, 2013) and negative emotion (Chang et al., 2015). Our analyses applied a multilevel mediation framework to i) identify immediate and lasting effects of cognitive regulation (i.e. up-regulating versus down-regulating) on experienced negative affect, ad effectiveness, and ad self-relevance, ii) identify effects of cognitive regulation on brain responses associated with emotion and value, and iii) ask whether cognitive regulation of emotion- and value-related brain responses formally mediated changes in experienced negative affect, ad effectiveness, and ad self-relevance.

Method

Participants

Participants were 60 adults recruited and screened to confirm that they typically drank alcohol at least two to three times a month, were right-handed, could read and speak fluently in English, had normal or corrected-to-normal vision, had never been diagnosed with a psychiatric or neurological disorder, were not currently using psychiatric medication or legally prohibited drugs, were not currently pregnant or

breastfeeding, and had no conditions that contraindicated MRI. All procedures were approved by the Institutional Review Board at the University of Pennsylvania. Two participants were excluded from analysis due to data corruption (one due to severe MRI dropout, and one due to excessive head motion) and one participant's data was lost due to a scanner system failure, leaving a final sample of 57 (32F) adults (mean age=22.9, SD=2.97).

Image acquisition

Data were acquired on a 3T Siemens Prisma scanner with a 64-channel head/neck array. Structural volumes were acquired using a high-resolution T1-weighted axial MPRAGE sequence yielding 160 slices with a 0.9 by 0.9 by 1.0mm voxel size. Functional volumes were acquired using a T2*-weighted image sequence with a repetition time (TR) of 1000ms, an echo time (TE) of 32ms, a flip angle of 60°, and a 20cm FOV consisting of 56 with 2.5mm thickness acquired at a negative 30° tilt to the AC-PC axis, with a 2.5mm isotropic voxel size and a multiband factor of four. Finally, we collected an in-plane structural T2-weighted image consisting of 176 axial slices with 1mm thickness and 1mm isotropic voxel size to implement a two-stage coregistration procedure between functional and anatomical images.

Design

Scanner cognitive regulation task. Participants completed a cognitive regulation task across four functional runs in which they were asked to view static visual anti-binge drinking ads and either respond naturally or try to deliberately control their response. Before the scan, participants completed an experimenter-guided training module that provided examples of strategy implementation and opportunities to practice. For the

emotion-focused strategy, participants were told to respond as they normally would (look naturally), or to think about the situations depicted and information conveyed within the ad in a way that makes them feel more negative emotion (up-regulate) or less negative emotion (down-regulate). For example, to down-regulate negative emotion participants could imagine that the depicted imagery was staged or edited, or otherwise not as bad as it appeared. For the persuasion-focused strategy, participants were told to respond as they normally would (look naturally), to think about the situations depicted and information conveyed by focusing on what is persuasive (up-regulate) or by focusing on what is not persuasive (down-regulate). For example, to up-regulate persuasiveness, participants could focus on an aspect of the argument presented that was particularly compelling to them. For both strategies, participants were instructed to pay attention to each ad and to change the way they thought about the ads, but not to look away or distract themselves.

One half of the scanner task (two consecutive runs) was devoted to an emotion-focused cognitive regulation strategy (with down-regulate emotion trials, look naturally trials, and up-regulate emotion trials intermixed), and the other half (two consecutive runs) was devoted to a persuasion-focused cognitive regulation strategy (with down-regulate persuasion trials, look naturally trials, and up-regulate persuasion trials intermixed). The order of the two strategies was counterbalanced across participants. The entire task consisted of 90 trials: 30 down-regulate trials, 30 look naturally trials, and 30 up-regulate trials. The trial sequence, consisting of a 2s instructional cue, 8s ad presentation, two consecutive 4s rating periods, and a 3-7s inter-trial interval (ITI), is represented in Figure 1. Each participant viewed 90 ads that were counterbalanced to

experimental condition across participants, ensuring equal rates of ad allocation across condition (down-regulate, look naturally, and up-regulate across emotion-focused and persuasion-focused strategies). On each trial, participants rated how negative they currently felt (1: *not at all* to 5: *extremely*) and how effective they found the ad to be (1: *not at all* to 5: *extremely*). Ratings of negative affect and perceived message effectiveness gave an immediate measure of the effects of the instructed strategies on the aspects of psychological experience that they were directly targeted by the cognitive strategies. Stimuli were presented with PsychoPy v1.9, and participants made behavioral responses on a five-button response pad.

The ads used in this task were static images that were drawn from a search of online anti-binge drinking campaign websites and social media pages. This resulted in a set of 129 distinct ads, each of which included text and images and was designed to communicate the risks of excessive alcohol consumption. After collecting these ads, we conducted a norming study in which each ad was rated by at least 20 independent coders who viewed an ad for at least 8 seconds, yielding normative ratings of each ad. We used these ratings to select a set of 90 ads that were highest in normative ratings of persuasiveness and negative affect. The average normative negative affect elicited by these 90 stimuli was 2.4 (SD=0.3), and average normative persuasiveness was 2.6 (SD=0.2) (on a scale from 1: *not at all* to 5: *extremely*).

Post-scan re-exposure task. After the scan (about one hour after the reappraisal task in which participants viewed the ads in the fMRI scanner), participants completed a self-paced ad re-exposure task (see Figure 1) in which they were re-exposed to 45 of the 90 ads they viewed within the scanner task and asked to rate their current negative

affect, the effectiveness of the ad, and the self-relevance of the ad (1: *not at all* to 5: *extremely*). Post-scan ratings of message self-relevance were included to give a measure of the extent to which the effects of the instructed strategies generalize in an enduring way to another important component of how messages are subjectively experienced.

Analysis

Preprocessing and general linear model (GLM). Data preprocessing incorporated tools from SPM8, AFNI and FSL, and consisted of despiking, slice-time correction, realignment, coregistration of functional and structural images, and normalization to the standard Montreal Neurological Institute (MNI) brain by segmentation of the structural image. Normalized images were smoothed with an 8mm Gaussian kernel.

We constructed first-level (individual participant) GLMs in SPM8 to estimate primary contrasts of interest. Cue (2s), stimulus (8s), and response (8s) periods of each trial were modeled as boxcar functions convolved with the canonical hemodynamic response function. A single regressor was entered for all cue periods, and another regressor for all response periods. Six separate regressors were entered for stimulus presentation periods within: down-regulate emotion, look naturally, and up-regulate emotion (within the emotion-focused task runs), and down-regulate persuasion, look naturally, and up-regulate persuasion (within the persuasion-focused task runs). Six rigid-body motion parameters and a high pass temporal filter for 128 seconds were added as regressors of no interest.

We also constructed a single-trial GLM to quantify trial-level estimates of brain activity (Koyama et al., 2003; Rissman, Gazzaley, & D'Esposito, 2004), to use in

multilevel predictive modelling. Within this GLM, each stimulus (ad-viewing) period of the task was modeled as a separate boxcar function convolved with the canonical hemodynamic response, generating separate estimates of brain activity (relative to implicit baseline) for each ad-viewing period, for each participant. Regressors for cue and behavioral response periods, six rigid-body motion parameters, and a high-pass filter for 128 seconds were included as regressors of no interest.

We implemented second-level (group) random-effects analyses in NeuroElf v1.1 (neuroelf.net). Our primary analysis consisted of a whole-brain random-effects analysis of variance (ANOVA) with trial type (down-regulate, look naturally, up-regulate) and strategy (emotion-focused, persuasion-focused) as within-person factors. To threshold whole-brain results, we applied parametric cluster-extent thresholding using Monte Carlo simulation to achieve a whole-brain familywise error rate (FWER) corrected p-value of $< .05$, with a primary threshold of $p = .001$ and smoothness parameters (14.4 to 15.7mm) estimated from the residuals of each statistical map. This yielded a minimum number of contiguous voxels, k, from 363 to 427 for individual maps. For search analyses within ROIs, we applied small-volume correction to achieve a corrected $p < .05$, using Gaussian Random Field theory to estimate the number of independent resolution elements in each ROI.

Regions and patterns of interest. We defined regions and patterns of interest (ROIs and POIs) in order to estimate brain activity associated with core psychological processes of interest. Coordinates refer to ROI center of mass in MNI space. Given the focus of past work about the neural predictors of persuasion-induced health behavior change on vmPFC (Falk & Scholz, 2018), a vmPFC ROI was defined on the basis of

our whole-brain random effects ANOVA; this analysis isolates the vmPFC cluster that showed an omnibus main effect of trial type (at p = .001) (-2, 32, -11; 16551 mm^3). (For analyses with a region of vmPFC previously identified as being involved in positive reappraisal of negative stimuli, see Supplementary Materials). Because distributed patterns of activity can provide behaviorally-relevant information about brain states beyond activity in isolated ROIs (Chang et al., 2015; Wager et al, 2013), we also defined a pattern of interest indexing neural processes related to subjective value (Bartra, Mcguire, & Kable, 2013) and another pattern of interest indexing neural processes related to negative emotion (Chang et al., 2015).

A pattern of interest is a generalization of the concept of an ROI wherein voxels are assigned continuous weights rather than a binary assignment of being included in an ROI or not. To index distributed neural processes related to valuation, we used results from a meta-analysis of 206 studies identifying neural regions responding to value (Bartra et al., 2013). Specifically, we used the contrast of positive subjective value effects over negative subjective value effects (i.e., the unthresholded *t* map used to create Figure 3D from Bartra et al., 2013), which resulted in a pattern wherein the value for each voxel reflects the extent to which nearby activity is reliably associated in the existing literature as positively tracking (more so than negatively tracking) with subjective value. To index distributed neural processes related to negative emotional reactivity, we used a whole-brain negative emotion pattern, developed with regularized regression, that reliably tracks with ratings of negative emotion elicited by aversive images (Chang et al., 2015). The negative emotion and valuation patterns we focused on are spatially distinct and were only weakly correlated in their expression from trial to

trial (1% to 3% shared variance), suggesting that they provided non-redundant information about global brain responses. Moreover, the patterns were also only weakly or moderately correlated with activity in the subregion of vmPFC that constituted our omnibus effect ROI (vmPFC activity showed approximately 0% shared variance with negative emotion pattern expression and 8% shared variance with valuation pattern expression) supporting the notion that these distributed patterns carry information different than what is captured by isolated ROIs.

Pattern expression analyses. We conducted pattern expression analyses to test whether expression of our whole-brain patterns of interest were i) influenced by cognitive regulation and ii) predictive of immediate and lasting behavioral responses. In order to calculate the extent to which trial-level beta images expressed a pattern of interest, we treated the pattern as a vector of weights and calculated the dot product between this vector and each vectorized trial-level brain activation image, yielding a scalar value reflecting the extent to which the pattern of interest was expressed on each trial, for each participant.

Multilevel predictive modelling. We used R (cran.r-project.org; ver 3.3.1), Stan (mc-stan.org; rstan ver 2.16.2), and the brms package (Bayesian Regression Models using Stan ver 2.1.0) to fit hierarchical Bayesian regression models that estimated the extent to which brain activity within our ROIs and expression of our patterns of interest i) differed according to experimental conditions and ii) were predictive of ad-to-ad differences in reported negative emotion, ad effectiveness, and self-relevance. We also fit hierarchical Bayesian structural equation models to test whether effects of experimental conditions on emotional and attitudinal outcomes were mediated by

changes in brain activity (see Figure 1B). For models comparing different kinds of predictors (e.g., normative ratings of the ads, effects of cognitive regulation, and brain responses), variables were standardized and (for variables that varied within-person) person-mean centered, yielding standardized beta coefficients indicating the average magnitude of the within-person relationship between the predictor variable and the outcome variable. Models incorporated variance and covariance parameters allowing for model intercepts and slopes to vary by person and by stimulus. We used posterior means and 95% Bayesian credibility intervals (posterior highest density intervals) to estimate the plausible range of values that a given relationship could take in light of the observed data.

Because weakly informative priors centered at zero yield results that closely correspond with traditional maximum likelihood estimates (but regularize extreme values toward zero), we used weakly informative priors on beta coefficients, variance parameters, and covariance parameters. Specifically, we used a zero-centered t distribution with scale parameter 10 and 3 degrees of freedom for beta coefficients, a positive half-t distribution with scale parameter 10 and 3 degrees of freedom for standard deviations, and an LKJ distribution with regularization parameter 1 for correlations between person-level intercepts and slopes (Stan Development Team, 2016). Models were estimated with Markov Chain Monte Carlo Sampling, running four parallel chains for 1000 iterations each (the first 500 warm-up samples for each chain were discarded). This number of iterations proved sufficient for convergence in that the Gelman-Rubin diagnostic reached a value between 0.95 and 1.05 for all parameters (Gelman & Rubin, 1992). In comparison to maximum likelihood based approaches to

multilevel modelling, this approach offers: posterior inference, more accurate estimation of hierarchical variance and covariance parameters, better rates of convergence, and diagnostics for assessing the validity of the sampler-based statistical inferences (Stan Development Team, 2016).

Results

Cognitive regulation of persuasive messages evoked goal-congruent changes in negative affect and perceived effectiveness

First we asked whether cognitive regulation (i.e., reappraising the meaning of a persuasive ad in an attempt to change its emotional or persuasive impact) had immediate effects on negative affect (related to the depicted consequences of drinking) and perceived ad effectiveness. Collapsing across the emotion-focused and persuasion-focused strategies, relative to natural responding, participants reported increased negative affect when up-regulating, b=.24, 95%CI[.15, .33], and decreased negative affect when down-regulating, b=.21, 95%CI[.13, .29]. However, there was also an interaction with strategy, b=.39, 95%CI[.27, .51], such that the emotion-focused strategy more strongly modulated negative affect (i.e., up-regulate versus down-regulate), b=.64, 95%CI[.50, .77], than did the persuasion-focused strategy, b=.25, 95%CI[.12,.38] (see Figure 2A).

Turning to ratings of ad effectiveness, participants reported increased ad effectiveness when up-regulating their response to the ad, b=.25, 95%CI[.16, .34], and decreased ad effectiveness when down-regulating, b=.31, 95%CI[.23,.40]. Specifically, there was an effect of cognitive regulation (up- versus down-regulation) for both the emotion-focused strategy, b=.53,95%CI[.43, .62], and the persuasion-focused strategy,

b=.60,95%CI[.50, .70], with no interaction indicating a clear difference between these two effects, b=.07, 95%CI[-.06, .21]. There was also main effect of strategy block such that participants gave higher effectiveness ratings within blocks where they applied the persuasion-focused strategy, b=-.15, 95%CI[-.21,-.10] (see Figure 2A). Overall, this pattern of results indicates that the emotion-focused strategy modulated negative affect somewhat more than the persuasion-focused strategy did, but the two strategies modulated perceived ad effectiveness to a comparable extent.

Cognitive regulation of persuasive messages evoked lasting change in perceived message self-relevance

Outside of the scanner, participants completed a re-exposure task in which they viewed 45 of the 90 ads they viewed in the scanner (approximately 1 hour after having seen them previously) and rated negative affect, perceived effectiveness, and self-relevance for each ad. Analyses of these data revealed an effect of trial type on self-relevance, b=.13, 95%CI[.02, .23], with no interaction by strategy, b=.07, 95%CI[-.13, .28], indicating that ads that had been previously shown in the up-regulation condition were perceived as more self-relevant than ads previously shown in the down-regulation condition after a one hour delay (see Figure 2B). There wasn't a clear lasting effect of trial type on ratings of perceived effectiveness, b=.06, 95%CI[-.02, .14], or negative affect, b=.04, 95%CI[-.06, .15], at re-exposure (90% and 77% of the posterior densities for these effects were above zero).

Cognitive regulation evoked goal-congruent modulation of vmPFC, and whole-brain patterns associated with emotion and value

Next we turned to the fMRI data, focusing on the period of ad presentation during which participants were exposed to the messages (and either responded naturally or implemented cognitive regulation). To identify regions of the brain showing differential activity across experimental task conditions, we fit a 3 (trial type: down-regulate, look naturally, up-regulate) by 2(strategy: emotion-focused, persuasion-focused) whole-brain ANOVA. Several regions showed an omnibus main effect of trial type (down-regulate, look naturally, up-regulate), including bilateral ventrolateral prefrontal cortex, dorsolateral prefrontal cortex, dorsomedial prefrontal cortex (pre-supplementary motor area), ventromedial prefrontal cortex (shown in Figure 3A), and posterior parietal cortex (FWE $p < .05$), all of which have been previously implicated in controlled processing and emotion regulation. Across these regions, a pattern was apparent whereby activity was higher for up-regulate trials versus look naturally trials, $b = .08$, 95%CI [.06, .11], and also higher for down-regulate trials versus look naturally trials, $b = .06$, 95%CI[.03, .08], consistent with a role for these regions in implementing cognitive regulation when either up- or down-regulating (see Supplementary Figure S1). Two other regions not typically implicated in implementing cognitive regulation of emotion (Buhle et al., 2014) were also identified as showing an omnibus effect of trial – ventromedial prefrontal cortex and posterior cingulate. Within the vmPFC cluster, activity was higher for up-regulate trials than for down-regulate trials, $b=.02$, 95%CI[.01, .04], consistent with goal-congruent modulation of activity in this brain region (see Figure 3A). We also observed a similar pattern in a cluster within posterior cingulate (shown in Figure 3A).

In parallel to our univariate analyses, we conducted multivariate pattern expression analyses that leveraged two distinct whole-brain patterns of interest: a

negative emotion pattern that is highly predictive of negative emotion elicited by aversive images (Chang et al., 2015), and a valuation pattern meta-analytically defined as tracking with subjective value (Bartra et al., 2013). Here, the negative affect participants reported was related to the depictions of negative consequences of drinking within the anti-binge drinking ads. Consistent with goal-congruent modulation of negative affect- and value-related global brain activity, there was a main effect of trial type such that expression was higher for up-regulate trials than for down-regulate trials for both the negative emotion pattern, b=.08, 95%CI[.02, .13], and the valuation pattern, b=.26, 95%CI[.20, .33] (see Figure 3B, 3C). Notably, valuation pattern expression was modulated such that expression was lowest on down-regulate trials, but comparably high on look naturally and up-regulate trials. On the other hand, negative emotion pattern expression was modulated such that expression was highest on up-regulate trials but comparably low on down-regulate and look naturally trials (see Figure 3B, 3C).

vmPFC activity, valuation pattern expression, and negative emotion pattern expression predicted ratings of negative affect, perceived effectiveness, and message self-relevance

The above analyses revealed that cognitive regulation brought about goal-congruent changes in brain activity associated with negative emotion and value, but they did not speak to the relevance of these changes for subjective emotional and evaluative responses to the persuasive ads. To address this, we asked whether trial-to-trial differences in ROI activity and pattern expression could predict immediate and lasting change in emotion and message appraisals. We built multilevel models using these brain variables to predict negative affect, ad effectiveness, and self-relevance.

We first considered the behavioral ratings made in the scanner, immediately after viewing the ad and implementing the instructed strategy. We found that in-scanner negative affect ratings were predicted by trial-to-trial differences in vmPFC activity, b=.07, 95%CI[.03, .10], and negative emotion pattern expression, b=.09, 95%CI[.05,.13]. To ask whether these brain variables tracked with negative affect above and beyond an association with ratings of ad effectiveness, we ran the same models while additionally controlling for ratings of negative affect. When doing so, the predictive effect of negative emotion pattern expression held, b=.05, 95%CI[.01,.08], but the vmPFC predictive effect did not, b=.02, 95%CI[-.02,.05], suggesting that vmPFC activity predicted negative affect largely to the extent that negative affect shared variance with ratings of ad effectiveness. In-scanner ad effectiveness ratings were strongly predicted by vmPFC activity, b=.11, 95%CI[.07, .14], negative emotion pattern expression, b=.09, 95%CI[.06, .12], and valuation pattern expression, b=.13, 95%CI[.10, .16], and these effects held after additionally controlling for in-scanner ratings of negative affect.

Finally, we asked whether brain responses measured during exposure to the ads within the scanner could predict self-relevance ratings made one hour later, when participants were re-exposed to the ads (without a cue prompting them to think about the ad in a particular way). Here we found that post-scanner ratings of ad self-relevance were strongly predicted by vmPFC activity, b=.09, 95%CI[.03, .14], negative emotion pattern expression, b= .05, 95%CI[.01, .09], and valuation pattern expression, b=.05, 95%CI[.01, .10].

Change in negative affect, perceived effectiveness, and message self-relevance was mediated by cognitive regulation of brain responses associated with emotion and value

Our previous analyses assessed relationships between our experimental manipulation of cognitive regulation and brain activity, and between brain activity and behavioral ratings, but they did not jointly test a mediation model assessing whether a given pattern of brain activity links an experimental manipulation with a resulting rating outcome, either for immediate ratings or for delayed ratings made at re- exposure one hour later. To do so, we fit multilevel structural equation models testing whether our a priori regions and patterns of interest mediated the within-person effect of cognitive regulation (i.e., trial type: up-regulate versus down-regulate) on subjective experience (i.e., immediate ratings of negative affect, and ad effectiveness, and delayed ratings of self-relevance). We tested negative emotion pattern expression, valuation pattern expression, and vmPFC as parallel mediators of within-person effects of cognitive regulation, consistent with a model whereby these brain variables provide independent contributions in constructing emotions and attitudes. As shown in Figure 3, these models indicated: 1) the immediate effect of the experimental manipulation of cognitive regulation (up- versus down-regulate) on ratings of negative affect was mediated by negative emotion pattern expression (indirect path = 0.0026, 95%CI[.0017, .0047]) and vmPFC activity (indirect path = 0.0016, 95%CI[.0010, .0037]) in parallel, 2) the immediate effect of cognitive regulation on perceived ad effectiveness was mediated by negative emotion pattern expression (indirect path = 0.0022, 95%CI[.0006, .0044]), valuation pattern expression (indirect path = 0.0082, 95%CI[.0040, .0132]), and vmPFC

activity (indirect path = 0.0023, 95%CI[.0001, .0049]) in parallel, and 3) the lasting effect of cognitive regulation on delayed ratings of self-relevance was mediated by vmPFC activity (indirect path = 0.0023, 95%CI[.0006, .0047]). There was also marginal evidence that negative emotion pattern expression mediated the lasting effect of cognitive regulation on self-relevance in parallel to the vmPFC mediation pathway (97.4% of the posterior density for the mediation effect was above zero). Overall, these results indicate that brain activity associated with emotion and value formally mediated the effect of cognitive regulation of persuasive anti-binge drinking messages on resulting emotions and attitudes, including an experimentally induced increase in the perceived self-relevance of ads apparent after a one-hour delay.

Discussion

People can think about persuasive messages in ways that undercut or enhance their impact, an important factor underlying whether attempts to persuade succeed or fail. We first demonstrated a causal path from experimentally manipulated psychological goals to persuasive outcomes, and then performed a test of the brain mechanisms underlying these cognitively evoked changes. We focused on responses to ads communicating risks of excessive alcohol consumption, a significant public health problem, where prevention efforts are especially needed (NIAAA, 2015). Specifically, we examined neural pathways by which goals to cognitively enhance versus dampen responses to a persuasive message could shape resulting emotions and attitudes.

We found that cognitively enhancing versus diminishing responses to ads evoked robust change in negative affect and perceived ad effectiveness, and also evoked a difference in perceived ad self-relevance that was apparent after a one-hour delay.

Using a within-person mediation approach, we found that immediate effects of cognitive regulation on negative affect and perceived effectiveness, as well as delayed effects on self-relevance, were mediated by cognitive modulation of emotion-, and value-related brain responses. Thus, the experimentally-induced change in the perceived self-relevance of ads apparent after a one-hour delay was mediated by modulation of vmPFC activity. Overall, these data suggest that deliberately controlling one's appraisal of a persuasive appeal can modulate brain responses associated with emotion and value that in turn shape the impact of the appeal in durable manner.

Implications for psychological and neural models of persuasion

Where current models highlight the role of brain valuation responses in predicting persuasion effects (Falk & Scholz, 2018), the results of this study extend this work in several ways. First, it is unclear from prior work whether valuation responses to persuasive messages reflect stable receiver-level or message-level differences, or whether they are flexibly constructed in a manner that is sensitive to goals and context. Our results provide evidence for the latter account by demonstrating that vmPFC activity can be modified in accordance with goals to reflect on an ad in a way that enhances versus dampens one's response. Similarly, although self-relevance plays a role in psychological models of persuasion (Petty & Cacioppo, 1990; Sherif & Hovland, 1961), it has typically been conceptualized as an input to the evaluation of a message, rather than an output of a goal-driven appraisal process. Here we show that perceived self-relevance of a message can be shaped by cognitive regulation, and identify a neural pathway – activity within vmPFC – that contributes to this effect. Finally, this study used a within-person mediation approach that allowed us to test functional pathways relating

cognitive goals to brain activity and persuasion effects that have not previously been examined.

Overall, the results of our analyses suggest that cognitive regulation can shape emotion- and value-related brain responses to a persuasive message, and these brain responses in turn shape affect and attitudes. In one result, valuation-related brain activity was comparably high when participants were instructed to look naturally at the persuasive messages as when they were instructed to cognitively up-regulate their response. There are at least two potential explanations for this finding that could be investigated in future work. First, people may tend to spontaneously appreciate the positive attributes of messages as a baseline state. Additionally, it could also be that, on average, people are less effective in cognitively up-regulating neural valuation activity than in cognitively down-regulating neural valuation activity. Future work in this area could extend our findings by seeking to directly estimate the causal impact of brain responses on affect and attitudes with brain stimulation techniques that afford direct experimental control of brain activity.

Implications for persuasive messaging interventions

A common observation in the health communication literature is that people who are most at risk for disease are often least receptive to messages that communicate these risks (Resnicow & McMaster, 2012), because they disproportionately respond to persuasive messages by attempting to counter-argue or refute them (Dillard & Shen, 2005). In line with this idea, prior work suggests that activity within regions of the brain associated with cognitive control can relate to anti-drug messaging effects in individuals rated as high in risk for cannabis use, and hence likely to engage in strategies to

derogate the messages' effects (e.g., counter-arguing, Weber, et al., 2014), and that smokers showing activity in regions of dorsolateral prefrontal cortex implicated in counter-arguing later use more negative, deliberative language in describing the messages (Liu et al, under revision). Here, participants engaged in a process analogous to counter-arguing in which they cognitively regulated their response to a persuasive ad to either enhance or diminish its effects. Broadly, the impact of any persuasive message is a function of the meaning that a person derives from the message, beyond the objectively measurable properties of the message per se (Bruner, 1957; Carey, 1989; Griffin & Ross, 1991).

With this framework in mind, a poor response to a persuasive message (e.g., a recipient shows no change in attitudes or reactance) could be caused by any of 1) low bottom-up reactivity (e.g., an ad that is normatively uncompelling), 2) a successfully enacted goal to cognitively diminish reactivity (e.g., the recipient reflects on the ad in a way that diminishes its value or self-relevance), or 3) an unsuccessfully enacted goal to cognitively enhance reactivity (e.g., the recipient reflects on the ad in a way intended to increases its value or self-relevance but fails to do so successfully). Drawing from research in persuasion that identifies factors that influence how deeply arguments are processed (Petty & Cacioppo, 1986), as well as models of the neural mechanisms predicting decisions to enact control of emotion (Doré, Weber, & Ochsner, 2017; Shenhav, Botvinick, & Cohen, 2013) future studies could focus on psychological and neural factors that influence decisions to deliberately reflect on the meaning of a persuasive appeal in order to change its impact. In particular, the position that decisions to deliberately control one's response to a persuasive stimulus can be treated as a

domain of value-based decision making may be a particularly powerful framework for understanding the brain bases of responses to persuasive messaging (Botvinick & Braver, 2014; Shenhav, Botvinick, & Cohen, 2013).

Related, although both positive and negative emotionally-evocative messages are known to impact decision-making (Gallagher & Updegraff, 2012; Rothman & Salovey, 1997), future work could ask how positive versus negative appeals differ in their malleability to cognitive regulation, or their tendencies to elicit specific kinds of cognitive regulation processes. For example, perhaps persuasive benefits of negatively valenced appeals may be undercut by counter-arguing processes they tend to evoke, but these counter-arguing tendencies could be mitigated with other kinds of interventions that promote specific kinds of cognition (Kang et al., 2018; Weber et al., 2014). Moreover, future work could ask whether specific populations (e.g., people with substance use disorders) show differences in bottom-up reactivity versus top-down regulation tendencies toward relevant and potentially-threatening health messages, and whether targeted cognitive or motivational interventions are able to normalize these responses (Falk et al., 2015; Kang et al., 2017).

Conclusion

Whether a persuasive message succeeds or fails depends not only on the message itself but on the goals and mindset of its recipient. Our data show that people can use cognitive strategies to diminish or enhance the effects of persuasive messages, and this ability relates to activity in brain systems associated with emotional reactivity, and integrative valuation. Specifically, brain responses associated with value mediated the effect of cognitive goals on perceived message self-relevance, suggesting a role for

the brain's valuation system in shaping responses to persuasive appeals in a manner that persists over time.

Acknowledgments

We thank Elizabeth Beard and Susan Hao for assisting with data collection. This research was supported by NIH New Innovator Award 1DP2DA03515601 (PI EBF), NIH/National Cancer Institute and FDA Center for Tobacco Products pilot grant (PIs EBF, BD) under TCORS grant P50CA179546 (PIs Robert Hornik and Caryn Lerman), the U.S. Army Research Laboratory including work under Cooperative Agreement Number W911NF-10-2-0022, and HopeLab. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Author Contributions

BD, NC, CS, MBO and EBF contributed to the study design. BD, NC, CS, and MBO collected the data under the supervision of EBF. BD analyzed data and drafted manuscript, with critical revisions from the other authors.

References

Bartra, O., McGuire, J.T., and 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.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. Annu Rev Psych. 66, 83-113.

Bruner, J.S. (1957). Going beyond the information given. Contemporary Approaches to Cognition 1, 119–160.

Buhle, J.T., Silvers, J.A., Wager, T.D., Lopez, R., Onyemekwu, C., Kober, H., Weber, J., and Ochsner, K.N. (2014). Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. Cereb. Cortex 24, 2981–2990.

Carey, J. (1989). Communication as culture: Essays on media and culture.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. J. Stat. Softw. 76.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. J. Pers. Soc. Psychol. 39, 752.

Chang, L.J., Gianaros, P.J., Manuck, S.B., Krishnan, A., and Wager, T.D. (2015). A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. PLoS Biol. 13, e1002180.

Clithero, J.A., and Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. Soc. Cogn. Affect. Neurosci. 9, 1289–1302.

Cooper, N., Bassett, D.S., and Falk, E.B. (2017). Coherent activity between brain regions that code for value is linked to the malleability of human behavior. Sci. Rep. 7, 43250.

Doré, B.P., Weber, J., and Ochsner, K.N. (2017). Neural Predictors of Decisions to Cognitively Control Emotion. J. Neurosci. 37, 2580–2588.

Falk, E., and Scholz, C. (2018). Persuasion, Influence, and Value: Perspectives from Communication and Social Neuroscience. Annu. Rev. Psychol. 69, 329–356.

Falk, E.B., Berkman, E.T., and Lieberman, M.D. (2012). From neural responses to population behavior: neural focus group predicts population-level media effects. Psychol. Sci. 23, 439–445.

Falk, E.B., O'Donnell, M.B., Cascio, C.N., Tinney, F., Kang, Y., Lieberman, M.D., Taylor, S.E., An, L., Resnicow, K., and Strecher, V.J. (2015). Self-affirmation alters the brain's response to health messages and subsequent behavior change. Proc. Natl. Acad. Sci. U. S. A. 112, 1977–1982.

Falk, E.B., O'Donnell, M.B., Tompson, S., Gonzalez, R., Dal Cin, S., Strecher, V., Cummings, K.M., and An, L. (2016). Functional brain imaging predicts public health campaign success. Soc. Cogn. Affect. Neurosci. 11, 204–214.

Fujita, K., Eyal, T., Chaiken, S., Trope, Y., and Liberman, N. (2008). Influencing Attitudes Toward Near and Distant Objects. J. Exp. Soc. Psychol. 227, 9044–9062.

Gallagher, K. M., & Updegraff, J. A. (2011). Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. Ann Beh Med. 43(1), 101-116.

Gelman, A., and Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Stat. Sci. 7, 457–472.

Griffin, D.W., and Ross, L. (1991). Subjective Construal, Social Inference, and Human Misunderstanding. In Advances in Experimental Social Psychology, M.P. Zanna, ed. (Academic Press), pp. 319–359.

Hovland, C.I., Janis, I.L., and Kelley, H.H. (1953). Communication and persuasion; psychological studies of opinion change (New Haven, CT, US: Yale University Press.).

Kang, Y., O'Donnell, M.B., Strecher, V.J., Taylor, S.E., Lieberman, M.D., and Falk, E.B. (2017). Self-Transcendent Values and Neural Responses to Threatening Health Messages. Psychosom. Med. 79, 379–387.

Koyama, T., McHaffie, J.G., Laurienti, P.J., and Coghill, R.C. (2003). The single-epoch fMRI design: validation of a simplified paradigm for the collection of subjective ratings. Neuroimage 19, 976–987.

Ochsner, K.N., Silvers, J.A., and 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–E24.

Pessoa, L. (2008). On the relationship between emotion and cognition. Nat Rev Neurosci. 9(2), 148.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. Ann Rev Neurosci. 35, 73-89.

Petty, R.E., and Cacioppo, J.T. (1986). Message Elaboration versus Peripheral Cues. In Communication and Persuasion: Central and Peripheral Routes to Attitude Change, R.E. Petty, and J.T. Cacioppo, eds. (New York, NY: Springer New York), pp. 141–172.

Petty, R.E., and Cacioppo, J.T. (1990). Involvement and persuasion: Tradition versus integration. Psychol. Bull. 107, 367–374.

Resnicow, K., and McMaster, F. (2012). Motivational Interviewing: moving from why to how with autonomy support. Int. J. Behav. Nutr. Phys. Act. 9, 19.

Rissman, J., Gazzaley, A., and D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage 23, 752–763.

Rothman, A. J., & Salovey, P. (1997). Shaping perceptions to motivate healthy behavior: the role of message framing. Psych Bull. 121(1), 3.

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

Sherif, M., and Hovland, C.I. (1961). Social judgment: Assimilation and contrast effects in communication and attitude change (Oxford, England: Yale Unievrsity Press).

Stan Development Team (2016). RStan: the R interface to Stan. R Package Version 2. 14. 1.

Weber, R., Huskey, R., Mangus, J. M., Westcott-Baker, A., & Turner, B. O. (2015). Neural predictors of message effectiveness during counterarguing in antidrug campaigns. Commun Monogr. 82(1), 4-30.

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. Commun. Monogr. 59, 329–349.

Figure Captions

Figure 1. A) Scanner cognitive regulation task and post-scanner ad re-exposure task. In the scanner task, participants saw a cue to either look naturally, to up-regulate their response, or to down-regulate their response (using an emotion-focused strategy or a persuasion-focused strategy). In the post-scan re-exposure task, participants viewed images they had seen in the scanner task and rated their current negative affect, perceived ad effectiveness, and ad self-relevance. **B) Multilevel mediation approach.** We used within-person mediation to ask whether a priori brain regions and patterns of interest (the mediator variables) could explain the effect of experimentally manipulated cognitive up- vs down-regulation (the predictor variable) on ratings of negative emotion, perceived ad effectiveness, and self-relevance (the outcome variables).

Figure 2. Behavioral results. Cognitive regulation evoked goal-congruent modulation of **A)** negative affect and perceived ad effectiveness ratings made in the scanner task, as well as **B)** change in ad self-relevance ratings made in the re-exposure task.

Figure 3 Effects of trial type on brain activity for A) vmPFC, a region identified as showing a whole-brain corrected omnibus main effect of trial type, and **B) expression of multivariate patterns associated with negative emotion and valuation.** (Plotted estimates reflect posterior means with 95%CI.)

Figure 4. Multilevel mediation. Effects of cognitive regulation – that is experimental instructions to up-regulate versus down-regulate one's response to an ad – on **A) in-scanner negative affect**, **B) in-scanner perceived effectiveness**, and **C) message self-relevance at re-exposure** were mediated by cognitively-driven change in brain responses associated with emotion and value. Path coefficients represent posterior means with 95% credible intervals; paths with 95% intervals that cross zero shown in lighter gray. Within line plots, black lines reflect overall group estimate, light gray lines reflect person-specific estimates. Visualization of posterior distributions reflect kernel density estimates of MCMC draws, with results from four parallel chains overlaid.