Intersubject correlations in reward and mentalizing brain circuits separately predict persuasiveness of two types of ISIS video propaganda

Michael S. Cohen[1*], Yuan Chang Leong[1], Keven Ruby[2], Robert A. Pape[2], Jean Decety[1,3*]

[1]Department of Psychology, University of Chicago, Chicago, IL
[2]Department of Political Science, University of Chicago, Chicago, IL
[3]Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL

*Corresponding authors:
Michael S. Cohen
Department of Psychology
University of Chicago
5848 S. University Ave.
Chicago, IL 60637
E-mail: mscohen@uchicago.edu

Jean Decety
Department of Psychology
University of Chicago
5848 S. University Ave.
Chicago, IL 60637
E-mail: decety@uchicago.edu

Abstract
The Islamist group ISIS has been particularly successful at recruiting Westerners as terrorists. A hypothesized explanation is their simultaneous use of two types of propaganda: Heroic narratives, emphasizing individual glory, alongside Social narratives, which emphasize oppression against Islamic communities. Functional MRI was used to measure brain responses to short ISIS propaganda videos distributed online. 4 Heroic and 4 Social videos were shown in the MRI scanner. Brain activity in reward circuitry (particularly ventral striatum), and in mentalizing regions such as dorsomedial prefrontal cortex (dmPFC), has been shown previously to predict preferences and persuasion respectively. Here, persuasiveness was measured using post-scan predictions of recruitment effectiveness. Inter-subject correlation (ISC) analyses were used to measure commonality of brain activity time courses across individuals. ISCs in ventral striatum predicted rated persuasiveness for Heroic videos, while ISCs in mentalizing and default networks, especially in dmPFC, predicted rated persuasiveness for Social videos. Other recent work has examined interactions between reward and mentalizing networks for persuasion in general, but the observed dissociation as a function of stimulus type is novel. These results provide evidence about neural mechanisms by which propaganda persuades prospective targets, while also supporting the hypothesized distinction between Heroic and Social narratives for ISIS propaganda.

Introduction

During the past decades, the Islamic State of Iraq and Syria (ISIS) has been uniquely successful at recruiting Western-born Muslims and converts to Islam to join their cause ([1], [2]). A central pillar of ISIS's recruitment strategy has been its propaganda, especially videos produced by the group in English and other Western European languages or featuring fighters from North America and Europe. Indeed, the vast majority of individuals indicted in the U.S. for supporting ISIS are American-born (61%), watched ISIS videos, and self-reported videos in their radicalization (85%) [1]. Understanding what makes ISIS videos effective in enhancing support for the group among diverse Western audiences advances the science of persuasion and political mobilization as well as contributes to countering the appeal of extremist groups.

One potential reason for ISIS's success in the West is the use of a "Heroic martyr" narrative alongside the "Social martyr" narrative more typical for terrorist propaganda. As research shows, ISIS Heroic narratives are analytically distinct from and occur more frequently in ISIS Western directed video propaganda than Social narratives [3]. Specifically, the Social martyr narrative features "good Muslims" who gain recognition for fulfilling their religious obligation to protect embattled communities based on strong ties to those communities. The Heroic martyr narrative, in contrast, casts its heroes as ordinary individuals who discover their true potential through extraordinary action, trading on tropes common to Western literature and action movies more than religious exegeses and sincerity of belief. The use of a traditional (the Social) and "westernized" (the Heroic) narrative has the potential to extend ISIS's reach among Muslims in the West who relate differently to Islam and life in the West.

Heroic narratives have been found to be particularly effective at appealing to those with weak ties to embattled Muslim communities while Social narratives appeal to those with stronger ties to such communities. A survey study found that American-born converts to Islam and Americans born into Muslim families in the U.S. rated Heroic videos as more persuasive than did immigrants from Muslim countries [4]. Additionally, when reviewing court records of actual ISIS members who were charged criminally in the United States for their terrorist activities, converts to Islam were scored as more likely to cite motivations in line with Heroic narratives while those born into Muslim families were more likely to indicate motivations in line with Social narratives [4]. Finally, an analysis of ISIS propaganda found that Western-directed propaganda from ISIS includes a mix of Social and Heroic narrative framings [4]. A key hypothesis following from these data is that ISIS may enhance its recruitment strategy success by choosing to rely on a mix of these narrative types specifically because they appeal to potential recruits in different ways.

One study using high-density electrophysiology (EEG) provided initial evidence that ISIS propaganda videos evoke distinct neural mechanisms based on whether they rely on a Heroic or Social narrative framing [5]. Most notably, Heroic videos were associated with increased frontal beta power, interpreted as reflecting greater personal relevance and positive

expectations. In contrast, Social videos are associated with reduced alpha power, reflecting greater engagement of attention for these videos, and greater frontal theta power, which could imply stronger emotion regulation. Together, these initial findings provided support for the theorized distinction between Heroic and Social videos [3]. The present work builds on these findings by using fMRI rather than EEG to measure brain activity and functional connectivity. Functional MRI has advantages relative to EEG in its ability to access the entire brain, including deep structures, and in providing more precise spatial localization of effects.

The same study also examined behavioral ratings of stimulus features and differences between individuals that modulate the effectiveness of ISIS propaganda videos. One feature typically shown to modulate persuasiveness is narrative transportation, i.e., the degree to which people can imagine themselves in the narrative and pay close attention to it (e.g., [6]). Prior work confirmed that narrative transportation predicted persuasiveness in an overlapping set of videos as that used in the present study, with a stronger relationship in men than in women [5]. Another factor that this prior work suggested may be relevant is participants' religiosity. Specifically, in a multivariate decoding analysis of EEG data, male participants who reported a higher degree of religious observance showed a greater difference in brain response for Heroic vs. Social videos than did male participants who were less religiously observant. Narrative transportation and religiosity measures were acquired for participants in the present study as well. Finally, justice sensitivity, a personality trait that affects how people react to injustices ([7], [8], [9], [10]), was examined for its possible relevance to rated persuasiveness of ISIS recruitment videos in the present study. This personality measure, some components of which are related to empathy [11], has been associated with extremist political beliefs in Western contexts. For instance, higher sensitivity to injustices where one is personally victimized, and lower sensitivity to injustices where others are victimized, were associated with greater support for both Donald Trump in the United States and for the far-right AfD party in Germany [12].

Neuroeconomics and neuroforecasting approaches

Past studies from the neuroeconomics and neuroforecasting literatures provide further motivation for the data analysis strategy used here. Activity in the reward/valuation system, particularly in the ventral striatum (VS)/nucleus accumbens (NAcc) and to a somewhat lesser extent in medial prefrontal cortex (mPFC), is greater in response to things that people choose, prefer, or like. This result is found in a variety of domains including facial attractiveness [13], consumer products [14], political candidates [15], newspaper articles [16], crowdfunding campaigns [17], and YouTube videos [18]. These reward responses can be modulated by the opinion of others [19] and social feedback [20]. Additionally, reward responses among a small sample of participants in the scanner can predict popularity of the same stimuli in a larger population (e.g., [16], [17], [18]).

There is some evidence that activity in the mentalizing circuit of the brain can also predict persuasiveness and attitude change. Note that ventral mPFC is often activated in mentalizing tasks (see, e.g., [21]) in addition to playing a key role in processing reward [22], but dorsal mPFC is more exclusively associated with social information processing. As has been noted in a prior review [23], studies of persuasion have diverged in emphasizing findings from dorsal vs. ventral regions of mPFC. Still, dorsomedial PFC (dmPFC) and other key brain regions that process social stimuli have been shown to be more active when processing more persuasive stimuli. In one study [24], text and video messages that participants rated as more persuasive, about a range of topic areas and across cultures, produced increased activity in dmPFC and bilaterally in posterior superior temporal sulcus (pSTS) and temporal pole (TP) relative to stimuli labeled as less persuasive. In addition, while many of the neuroforecasting studies that examine messaging effectiveness in the context of smoking cessation have focused on ventral mPFC ROIs, others have found critical effects in dorsal mPFC. This includes one study using messaging specifically tailored to participants' values and interests [25], and another that identified brain regions via an interaction between content and format that follows from the elaboration likelihood model of persuasion [26]. In both of the preceding studies, activity in dmPFC correlated with effectiveness of smoking cessation programs within the sample. Additionally, click-through rate of online anti-smoking ads in a large population was predicted by brain activity in a small scanned sample in brain regions including dmPFC, vmPFC, and L TPJ, as well as in posterior cingulate (PCC), a critical region for introspective processing, specifically in response to negatively-valenced smoking-relevant messaging [27]. Finally, in a different task domain, adolescents have been found to activate mentalizing regions when processing ratings from both parents and peers regarding the aesthetic value of art, and the degree to which these regions increase in activity corresponds to how much ratings tend to shift from a prescan baseline to match others' ratings [28].

Some work has also begun to examine the relationship between reward and mentalizing in persuasion. One study found that both reward and mentalizing circuits were activated when people changed their ratings of hypothetical video games in response to feedback from other people, but the degree of rating change was *negatively* correlated with functional connectivity between these networks [29]. This result suggests that the two networks make separable contributions to persuasion. More specifically, reward activity appeared to represent the value of the information regardless of its valence, while mentalizing activity tended to be stronger when ratings were impacted by negatively-valenced information. A different relationship between reward and mentalizing networks emerges from the value-based virality model, in the related domain of online content sharing. This work suggested an indirect effect by which increased activity in mentalizing and self-related processing regions of the brain leads to increased reward system activity, which then leads to greater content sharing [16]. Thus,

reward/valuation and mentalizing systems can work either independently or in concert depending on situational factors.

Inter-subject correlations

One important methodological difference between the current study and most prior work is that we examine inter-subject correlations (ISCs) and inter-subject functional connectivity (ISFC) in the brain. These methods more fully leverage the data available with naturalistic video stimuli, relative to more traditional analyses [30]. Specifically, these analyses focus on dynamic changes in brain activity over time, rather than measuring the mean amplitude of brain activity in response to a specific stimulus. Applying an ISC analysis to the time course of activity for an individual in a given brain region, versus the average time course for all other participants in that same region, identifies the degree to which processing in that region is stimulus-driven [31]. The response must also be shared across individuals, and not idiosyncratic, to drive an increase in ISCs estimated using this approach [32]. ISFC analyses are analogous to ISC analyses, but instead of examining the typicality of activity within *one* brain region, ISFC compares an individual's activity time course in one region with the group average of activity in *other* brain regions. This approach has been validated as improving the signal-to-noise ratio relative to functional connectivity measures computed within individuals [33]. It is specifically useful for measuring the degree to which nodes within and across networks are simultaneously activated in response to presented stimuli, while removing noise and intrinsic connectivity effects that are unrelated to the stimulus.

Some prior studies have related ISCs to specific stimulus features. One study related dynamic measures of valence and arousal with the time course of activity across the brain while watching a movie clip. This work found that ISCs in regions of the default/mentalizing network (e.g., dmPFC, pSTS, PCC), as well as ISCs in emotion processing regions including VS, were greater during points of the movie rated more negative in emotional valence, while ISCs in visual and dorsal attention regions (e.g., ventral visual cortex, frontal eye fields, intraparietal sulcus) were positively associated with momentary arousal ratings [34]. Another study found that higher rated persuasiveness of anti-alcohol public service announcements (PSAs) was associated with ISCs in a similar set of brain regions: mentalizing regions such as dmPFC and precuneus, as well as visual cortex, insula, inferior prefrontal gyrus, and supramarginal gyrus [35]. A study in which participants listened to political speeches, previously rated on how powerful they were perceived to be, found that more powerful speeches led to stronger ISCs in mPFC and in superior temporal gyrus [36]. Finally, meaningful personal narratives from the series "This I Believe" evoked greater ISCs in the mentalizing circuit, particularly dmPFC, TPJ, and precuneus, compared to a non-personal VCR instruction manual text control [37]. Thus, there is prior evidence suggesting that ISC analyses can be a valid approach to finding where in the brain commonality of signal fluctuations predicts communicative effectiveness.

In the analyses that follow, the strongest results emerge when ISCs in brain regions associated with reward and mentalizing are associated with persuasiveness ratings for each video. Specifically, persuasiveness of Heroic videos is modulated primarily by ISCs in a ventral striatum region of interest (ROI) [22], while persuasiveness of Social videos is modulated by mean ISC across a network of brain regions associated with mentalizing, particularly dmPFC [21]. These results suggest a double dissociation between stimulus type and the brain systems that lead to persuasiveness. Other results show aggregate differences in ISCs and ISFCs between Heroic and Social stimuli. These results together validate the hypothesized distinction between videos relying on a Heroic narrative and those relying on a Social narrative.

Results

Each participant was presented with four Heroic videos and four Social videos in the MRI scanner, with one video of each type presented per scan run in randomized order. Videos ranged in length from 34 to 63 seconds. The temporal window was shifted by 4 seconds for analysis, to account for delay in the hemodynamic response. Full neuroimaging data were acquired from 46 participants, but the post-scan questionnaire was only presented to the final 34 participants. In this questionnaire, participants re-watched and rated each video on measures of persuasiveness and narrative transportation. Data from the larger sample (n = 46) were averaged to compute group-level time courses to which each individual's data were compared, with the comparison individual always excluded from this average. Final results however are only reported from the subset (n = 34) who completed all behavioral measures.

Regression Analyses – Behavioral measures

Persuasiveness was used as the dependent measure for all regression analyses. It was operationalized as participants' ratings on a 1-7 scale to the following question: "This video would help the militant group recruit," collected after the functional MRI session about each video. All mixed-effects models were computed with the *lme4* and *lmerTest* packages in R using maximum likelihood estimation. An initial linear mixed effects model including clip type (Heroic or Social) and participant gender, with participants as a random effect, showed a main effect of clip type ($\beta = 0.368$, $t = 2.14$, $p = .033$) on persuasiveness, with Heroic stimuli rated as more persuasive, and no main effect of or interaction with participant gender (all $|t| < 1$).

Participants were also asked to rate each clip post-scan on three questions that together constituted a measure of narrative transportation. Prior to the scan session, each individual had responded to measures of Justice Sensitivity (4 subscales) [38] and Religiosity (3 subscales) [39]. In prior work [5], some relationships between behavioral variables and persuasiveness interacted with participant gender and/or with stimulus type. To determine which behavioral/demographic measures predicted rated persuasiveness, a series of linear mixed-effects models were run in which each variable and its interaction with participant gender and

stimulus type were separately regressed as fixed effects on persuasiveness, with participant modeled as a random effect. Gender was coded such that gender = 0 for males, gender = 1 for females, and gender = 0.5 for one non-binary participant. Stimulus types were coded such that Social = 0 and Heroic = 1. Across these models, only two behavioral variables showed a significant relationship with persuasiveness: Narrative Transportation and the Perpetrator subscale of Justice Sensitivity. A combined model that included both sets of variables yielded a lower AIC (701.47) than the model with only Narrative Transportation (713.17) or only Justice Sensitivity Perpetrator subscale (737.35); thus, this was chosen as the best behavioral model (Table 1).

Table 1. Behavioral model predicting rated persuasiveness (n = 34)

	β	t	p
(Intercept)	-0.101	-0.74	0.46
Participant Gender	-0.350	-1.92	0.059 ~
Stimulus Type	0.313	1.99	0.048
Justice Sensitivity (Perpetrator subscale)	0.125	0.96	0.34
Narrative Transportation	**0.323**	**3.17**	**0.002****
Participant Gender x Stimulus Type	**0.429**	**2.04**	**0.043***
Justice Sensitivity (Perpetrator subscale) x Participant Gender	0.344	1.77	0.081 ~
Justice Sensitivity (Perpetrator subscale) x Stimulus Type	**-0.416**	**-2.76**	**0.006****
Narrative Transportation x Participant Gender	-0.208	-1.45	0.15
Narrative Transportation x Stimulus Type	0.012	0.08	0.94
Justice Sensitivity (Perpetrator subscale) x Participant Gender x Stimulus Type	-0.070	-0.31	0.75
Narrative Transportation x Participant Gender x Stimulus Type	**0.485**	**2.32**	**0.021***

To better understand the interactions in behavioral effects shown in Table 1, we broke down these effects by gender and clip type. For male participants, there was a positive main effect of narrative transportation on persuasiveness ($\beta = 0.348$, $t = 3.14$, $p = .002$) and no

interaction between narrative transportation and clip type (β = 0.022, t = 0.14, p = .89). Thus, greater narrative transportation predicted higher persuasiveness ratings in males regardless of stimulus type. For the Justice Sensitivity Perpetrator subscale, men showed an interaction with clip type (β = -0.511, t = -2.75, p = .007) with no main effect (β = 0.143, t = 0.88, p = .38). Breaking the analysis down by clip type showed that men with a higher Justice Sensitivity Perpetrator score rated Heroic clips as less persuasive (β = -0.387, t = -2.41, p = .030), but no such relationship was apparent for Social clips (β = 0.148, t = 1.00, p = .32). In contrast, female participants showed an interaction between narrative transportation and clip type (β = 0.479, t = 3.55, p = .0005) with no main effect (β = 0.114, t = 1.21, p = .23). For the Justice Sensitivity Perpetrator subscale, women showed both an interaction with clip type (β = -0.357, t = -2.95, p = .004) and a main effect (β = 0.343, t = 3.39, p = .001). Breaking the analysis down by clip type showed that in women, higher narrative transportation predicted greater persuasiveness for Heroic clips (β = 0.643, t = 6.93, p < .0001) but not Social clips (β = 0.112, t = 1.00, p = 0.32). Higher Justice Sensitivity Perpetrator scores were associated in women with higher persuasiveness ratings for Social clips (β = 0.325, t = 2.80, p = .011) but not for Heroic clips (β = -0.016, t = -0.17, p = .86).

Predicting persuasion by reward-related brain activity

After identifying an optimal behavioral model, a key analysis of interest was the extent to which neural measures predicted additional variance in persuasiveness ratings. Brain regions associated with reward were of particular interest based on prior literature; ROIs were defined following [22]. Intersubject correlations (ISCs) were computed in right (R) and left (L) ventral striatum/nucleus accumbens (Vs/NAcc) ROIs, which are key nodes in the reward-sensitive circuit strongly associated with preferences (Figure 1A). When adding ISCs in Vs/NAcc, averaged across hemispheres, to the base model shown in Table 1, three different model specifications were tested: main effects only, interaction with stimulus type, and interactions with both stimulus type and participant gender. The model including an interaction with stimulus type yielded the lowest AIC value. An ANOVA comparing this model to the base model (the behavioral model shown in Table 1) showed that adding ISC values for Vs/NAcc predicted significantly more variance than the base model (χ^2 = 7.02, df = 2, p = .030). An interaction between Vs/NAcc ISC and stimulus type was observed on persuasiveness ratings (β = 0.236, t = 2.13, p = .034), but there was no main effect of Vs/NAcc ISC (β = -0.009, t = -0.12, p = .90). Separating the analyses by condition to examine simple effects found that Vs/NAcc ISC was a significant positive predictor of persuasiveness for Heroic videos (β = 1.310, t = 2.81, p = .006), but not for Social videos (β = -0.159, t = -0.36, p = .72). Figure 1B shows simple correlations between Vs/NAcc ISC and persuasiveness, which were similarly significant for Heroic clips (r = 0.21, p = .015) but not for Social clips (r = -0.03, p = .76). Thus, greater alignment in fluctuations

of Vs/NAcc activity across individuals predicted an increase in rated persuasiveness, specifically for Heroic clips.

This effect appears to be specific to VS/NAcc, rather than extending throughout the reward network. For ISCs averaged across R and L vmPFC, the model with the lowest AIC included an interaction with stimulus type, but this model predicted only marginally more variance than the base model (χ^2 = 4.89, df = 2, p = .087). The model showed a main effect of vmPFC ISC (β = 0.138, t = 2.13, p = .034) as well as a marginal interaction with stimulus type (β = -0.190, t = -1.85, p = .065). Simple effects did not show a reliable effect either in the Heroic condition (β = -0.337, t = -0.90, p = .37) nor in the Social condition (β = 0.546, t = 1.60, p = .113). Finally, when examining ISCs averaged across R and L ventral tegmental area (VTA) defined according to [40], adding only main effects to the base model produced a lower AIC value than models with any interaction terms, but this model still did not differ from the base model in predictive power (χ^2 = 0.01, df = 1, p = .94).

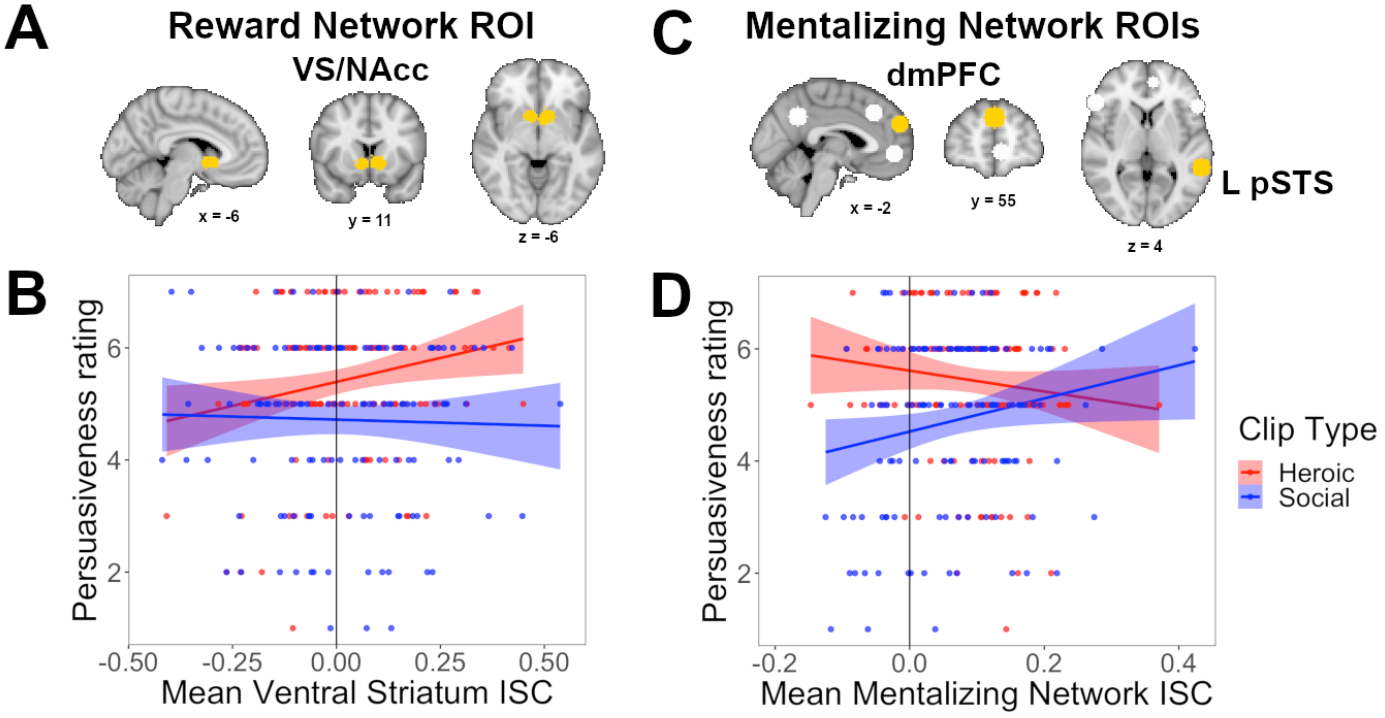

Figure 1. Bivariate relationships between ISC values and persuasiveness ratings for Heroic and Social stimuli. ISCs averaged across L and R hemispheres in ventral striatum/nucleus accumbens (VS/NAcc), displayed in yellow on panel (A), predict increased persuasiveness for Heroic stimuli but not for Social stimuli (B). Mean ISCs averaged across 14 ROIs in the mentalizing network, some of which are displayed in white or yellow on panel (C), particularly in dmPFC and L pSTS (highlighted in yellow on panel C), predict increased persuasiveness for Social stimuli but not for Heroic stimuli (D).

Predicting persuasion by mentalizing-related brain activity

Brain regions that belong to the mentalizing network have also been demonstrated in prior work to predict persuasiveness, and so form another network of a priori interest. The strongest result emerged when averaging ISCs across a set of 14 ROIs defined by activation peaks from a recent meta-analysis of mentalizing (Figure 1C) [21]. Here again, adding ISC values as well as an interaction term for stimulus type yielded the lowest AIC value. This model predicted significantly more variance than the base model (χ^2 = 13.88, df = 2, p = .001). A main effect of mentalizing network ISC was significant (β = 0.260, t = 3.63, p = .0003), as was an interaction between mentalizing network ISC and stimulus type (β = -0.338, t = -3.18, p = .002). Simple effect analyses showed that mentalizing network ISC was a significant predictor of persuasiveness for Social clips (β = 0.258, t = 3.08, p = .003), but not for Heroic clips (β = -0.059, t = -0.81, p = .42). Figure 1D shows the simple correlation between mean mentalizing network ISC and persuasiveness, which was significant for Social clips (r = 0.18, p = .038) but not for Heroic clips (r = -0.12, p = .18). Thus, the degree to which fluctuations in activity throughout the mentalizing network aligned across individuals predicted higher rated persuasiveness for Social clips.

These effects appear to be driven most strongly by ROIs in dorsomedial prefrontal cortex (dmPFC) and in left posterior superior temporal sulcus (L pSTS), highlighted in yellow in Figure 1C. A model including the main effect of ISC in dmPFC and its interaction with stimulus type predicted significantly more variance than the base model (χ^2 = 13.50, df = 2, p = .001). Here, the main effect of dmPFC ISC was significant (β = 0.222, t = 3.42, p = .0007), but with no interaction between dmPFC and stimulus type (β = -0.104, t = 0.99, p = .32). Breaking down the analysis by condition, to allow comparison with simple effects from other analyses, shows that the effect was significant for Social stimuli (β = 0.231, t = 2.96, p = .004) and was marginal in the same direction for Heroic stimuli (β = 0.123, t = 1.74, p = .085). For L pSTS, a model including both a main effect of ISC and an interaction with stimulus type again predicted significantly more variance than the base model (χ^2 = 7.78, df = 2, p = .020). This model showed both a significant main effect of L pSTS ISC (β = 0.183, t = 2.693, p = .008) and a significant interaction with stimulus type (β = -0.242, t = -2.38, p = .018). Simple effects confirmed a significant effect for Social stimuli (β = 0.184, t = 2.33, p = .021) but not for Heroic stimuli (β = -0.073, t = -1.04, p = .30). Adding ISCs in each of the other 12 individual ROIs from the meta-analysis [21] along with their interaction with stimulus type did not significantly improve predictive power relative to the base behavioral model, with uncorrected $p < .05$. Note that the effect in dmPFC survived FDR correction for multiple comparisons across all 14 regions (corrected p = .016), while the effect in L pSTS did not (corrected p = .14).

The relationship between ISCs in the mentalizing network and rated persuasiveness of Social clips was not limited to a particular definition of this network. When ISCs in the mentalizing network were computed by averaging across 9 ROIs defined via a different meta-

analysis [41], a model that added ISCs and their interaction with stimulus type to the behavioral model also significantly improved predictive power (χ^2 = 7.20, df = 2, p = .027). This model included a main effect of mentalizing network ISCs (β = 0.181, t = 2.61, p = .01) as well as an interaction with stimulus type (β = -0.235, t = -2.20, p = .029). Breaking the results down by stimulus type showed a significant effect for Social stimuli (β = 0.192, t = 2.39, p = .018) but not Heroic stimuli (β = -0.042, t = -0.55, p = .58). The same thing was observed using a mentalizing network defined by averaging across 12 mentalizing-related activations in a specific task that was intended as a more precise way of defining the mentalizing network than meta-analyses that combine data from varied mentalizing tasks [42]. Here again, a model with ISCs and their interaction with stimulus type significantly improved predictive power (χ^2 = 11.02, df = 2, p = .004). This model included a main effect of mentalizing network ISCs (β = 0.227, t = 3.16, p = .002) as well as an interaction with stimulus type (β = -0.317, t = -2.96, p = .003). Breaking the analysis down by stimulus type again showed a significant effect for Social stimuli (β = 0.212, t = 2.64, p = .009) but not Heroic stimuli (β = -0.073, t = -0.96, p = .34).

Relationship between reward and mentalizing on prediction

Another relevant question in light of prior work (e.g., [16]) is whether the relationships between persuasion and ISCs in reward and mentalizing circuits were independent from each other. A regression model including terms for both ISC in VS/NAcc (reward region) and mean ISC across the mentalizing network defined by [21], as well as interactions between each type of ISC and clip type, had lower AIC (687.77) than models that included only reward (698.45) or only mentalizing (691.58). As was the case when the two networks were analyzed separately, there was both a main effect of mentalizing network ISC (β = 0.260, t = 3.69, p = .0003) and an interaction between mentalizing network ISC and stimulus type (β = -0.351, t = -3.35, p = .0009), as well as an interaction between VS/NAcc ISC and stimulus type (β = 0.253, t = 2.34, p = .020), with no main effect of VS/NAcc ISC (β = -0.020, t = -0.30, p = .77). Breaking the analysis down by stimulus type showed, for Heroic clips, an effect of VS/NAcc ISC (β = 0.226, t = 2.88, p = .005) but no effect of mentalizing network ISC (β = -0.070, t = -0.96, p = .34). Social clips, by contrast, showed an effect of mentalizing network ISC (β = 0.257, t = 2.99, p = .003) but no effect of VS/NAcc ISC (β = -0.041, t = -0.50, p = .62). These results suggest that reward and mentalizing had independent and dissociable effects on persuasiveness.

Predicting persuasion by whole-brain activity

Finally, as an exploratory analysis, the degree to which ISCs in pre-defined parcels added predictive value to models predicting persuasiveness was examined in a set of 300 parcels assigned to 7 networks, defined according to a whole-brain parcellation [43]. When averaging ISCs across all parcels in a given network, and modeling both the main effect of ISC and its

interaction with stimulus type, only the Default network improved prediction of the model at an uncorrected p < .05 threshold. The model including both mean Default network ISC and its interaction with stimulus type predicted significantly more variance than the base model (χ^2 = 9.73, df = 2, *p* = .0077). However, this effect was just below the threshold for significance after FDR correction for multiple comparisons (p = .054). The model yielded both a main effect of ISC (β = 0.216, *t* = 3.15, *p* = .002) and an interaction with stimulus type (β = -0.210, *t* = -1.99, *p* = .047). Breaking down the model by stimulus type showed a significant effect for Social stimuli (β = 0.214, *t* = 2.69, *p* = .008) but not for Heroic stimuli (β = 0.013, *t* = 0.17, *p* = .86). Thus, greater alignment across individuals of fluctuations in Default network activity predicted higher rated persuasiveness for Social stimuli, similar to the results reported for the mentalizing network.

Figure 2A shows analogous results broken down by individual parcel. Specifically, the degree to which predictive value increased significantly relative to the base behavioral model when ISCs in individual parcels, and their interaction with stimulus condition, were included in the model, is shown for all parcels in which uncorrected p < .05. None of these effects were strong enough to survive FDR correction for multiple comparisons across all 300 parcels but are reported on an exploratory basis. Coefficients representing effects in the parcels showing an effect in Figure 2A were computed and plotted separately for Social (Figure 2B) and Heroic (Figure 2C) stimuli. Regions associated with mentalizing, including dmPFC, right inferior frontal gyrus, and precuneus, were among those that showed evidence of added predictive value, particularly for Social stimuli.

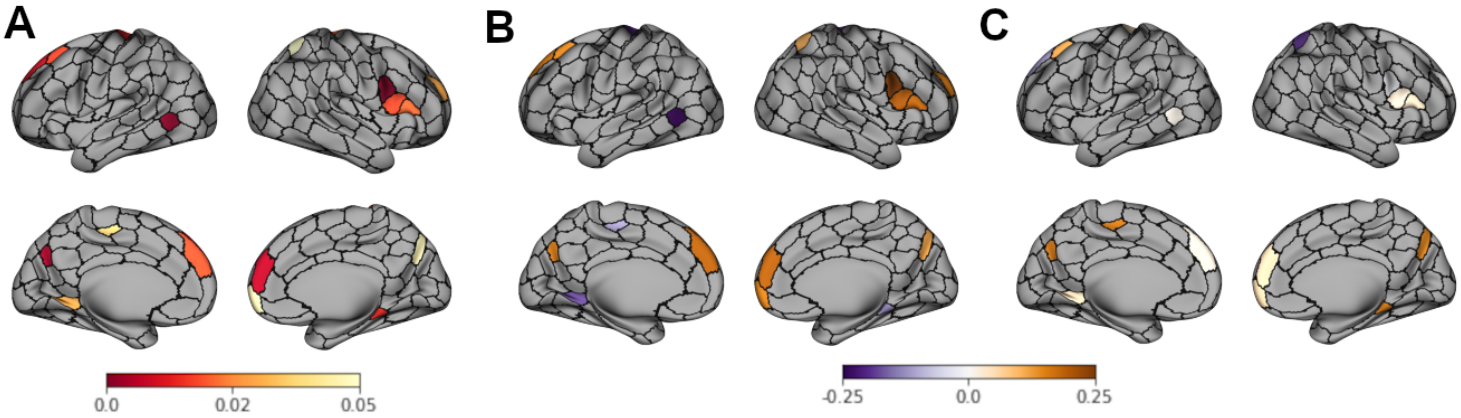

Figure 2. Exploratory parcel-based analysis showing where, across the whole brain, ISC values add to the predictive power of a behavioral model predicting persuasiveness. (A) p values (not corrected for multiple comparisons) reflecting increased predictive power when ISC in a particular region, and its interaction with stimulus condition, are added to the model. (B-C) β coefficients indicating the direction and strength of the relationship between ISC values and persuasiveness, in all parcels with p < .05 in panel A, for (B) Social and (C) Heroic stimuli.

Aggregate ISC/ISFC differences by clip condition

An alternative approach to examining these data is to compare how ISCs, as well as inter-subject functional connectivity (ISFC), differed between conditions when data from all stimuli of a given type are averaged together. These analyses are relevant to compare the fMRI results to our lab's prior EEG findings [5], as well as being a way to address how processing of these two types of stimuli differs on average. For ISC analyses, paired-samples permutation tests were used to compare each individual's mean ISC score for the 4 Heroic videos vs. the 4 Social videos. Mean ISCs were higher for Heroic videos when averaging ROIs across the mentalizing network, whether defined according to the set of activations from ref. [21] (p = .005), ref. [41] (p = .019), or ref. [42] (p = .006). However, no difference between conditions was apparent for the two specific mentalizing regions from [21] cited above as having most strongly predicted persuasiveness in Social videos: dmPFC (uncorrected p = .79) and L pSTS (uncorrected p = .58). With respect to reward ROIs, the VS/NAcc ROI showed no aggregate difference in ISCs by condition (uncorrected p = .72), nor did ROIs for vmPFC (uncorrected p = .43) or VTA (uncorrected p = .48).

A parcel-based analysis similar to that described above showed a number of parcels in which mean ISCs differed between Heroic and Social clips. The t-statistics for all regions with a significant difference in mean ISC by condition are plotted in Figure 3, thresholded using a paired permutation test statistic with FDR correction applied for multiple comparisons across all 300 parcels. Aggregating parcels by network, Heroic clips had higher ISCs in 4 of 7 networks, with FDR correction applied across the 7 networks: Visual (corrected p = .002), Dorsal Attention (corrected p = .023), Salience/Ventral Attention (corrected p = .034), and Default (corrected p = .003). There were no significant differences by condition in the Somatomotor (corrected p = .28), Limbic (corrected p = .22), or Control (corrected p = .09) networks.

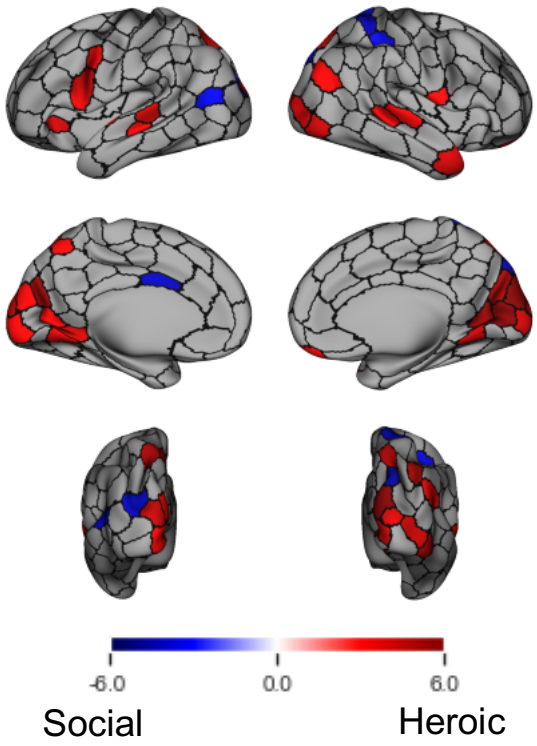

Figure 3. Parcels showing a difference in mean ISC by stimulus condition. Colors represent the t statistic for the difference in mean ISC value by condition. Map is thresholded based on parcels having corrected p < .05 following FDR correction on a paired permutation test.

Finally, ISFC analyses were used to determine whether the degree of coactivation between regions differed by condition. For this analysis, we aggregated edges either within a network or between two specific networks. After

computing p-values using permutation tests and applying FDR correction across 28 possible pairings, three pairings showed a significant difference between conditions (Figure 4A): Visual-Visual connections (corrected p = .029) and Visual-Dorsal Attention network connections (corrected p < .0001) were stronger for Social stimuli, while Visual-Default connections (corrected p = .028) were stronger for Heroic stimuli. As another way of representing these effects, raw connectivity matrices averaged across each of the 4 videos in each condition were submitted to a Network-Based Statistic (NBS) analysis [44]. This analysis allows for a visualization of individual connections that differ in strength between conditions, using assumptions about the spatial distribution of edges to correct for multiple comparisons. All edges connecting to Visual network nodes and showing a difference in connectivity by stimulus condition are plotted in Figure 4B and Figure 4C. The aggregate results shown in Figure 4A motivated limiting the plots to connections with Visual nodes. Similar to what was observed in the data aggregated by network, these plots confirm that Visual nodes were broadly more likely to be coactivated with Dorsal Attention nodes as well as with other Visual nodes in the Social condition (Figure 4B) relative to the Heroic condition (Figure 4C), while the reverse was true for connections between Visual and Default network nodes.

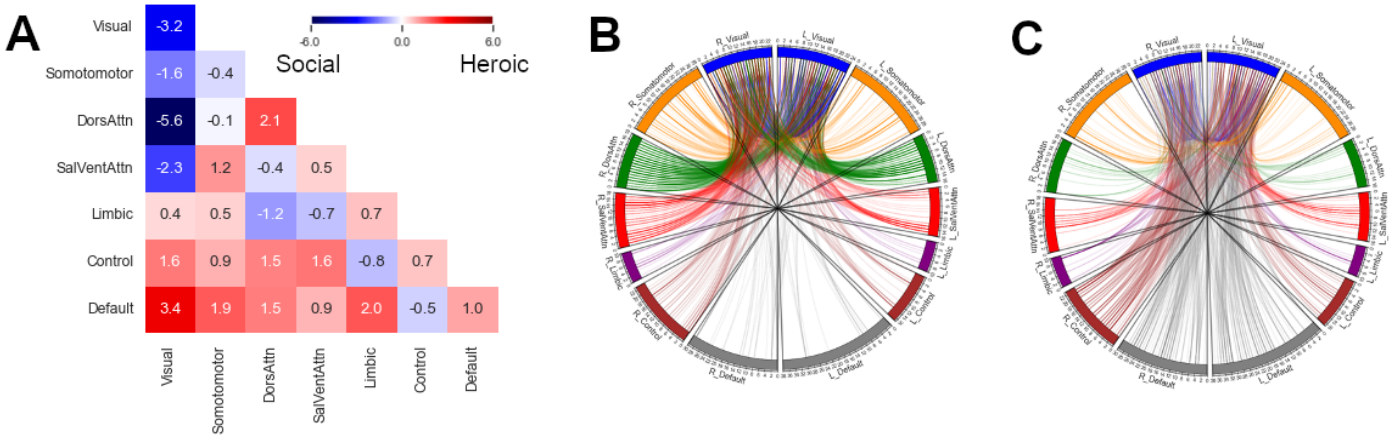

Figure 4. (A) Mean difference in ISFC values by condition when connectivity for all edges within a given network or pair of networks are averaged together. Colors represent the t statistic for the difference in mean ISFC value by condition. (B-C) Network-based statistic results showing all individual edges with a difference in connection strength by condition, with (B) Social > Heroic and (C) Heroic > Social.

Discussion

Greater ISCs in VS/NAcc predicted rated persuasiveness of Heroic videos but not of Social videos. Greater ISCs in brain regions associated with mentalizing, including most notably dmPFC, predicted rated persuasiveness of Social videos but largely not of Heroic videos. ISCs averaged across Default network parcels were similarly positively associated with

persuasiveness of Social videos but not Heroic videos. Heroic videos in the aggregate tended to produce larger ISCs than Social videos across visual, attention, and default network brain regions, potentially consistent with Heroic videos eliciting greater arousal than Social videos (cf., [34]). At the same time, ISFC analyses indicated that brain regions that process visual stimuli were more connected to other Visual regions and to Dorsal Attention regions when processing Social videos, but visual regions were more connected to Default network regions when processing Heroic videos. These effects suggest increased attentional focus directed to Social videos, potentially consistent with the finding from prior EEG work of reduced alpha power for Social videos [5]. Finally, behavioral results were consistent with prior work, at least in part, as Heroic videos tended to be rated as more persuasive, and narrative transportation also predicted persuasiveness independently from the effects computed from neural measures.

The functions of the specific brain regions in which ISCs were differentially associated with persuasiveness for Heroic and Social videos fits well with how those narrative categories have been conceptualized [3]. Heroic martyr narratives emphasize one's prospect of personal glory as an individual who achieves extraordinary feats for the group based on that person's personal capabilities. Activity in VS/NAcc has been associated in prior neuroforecasting literature with preferences in a variety of domains, all of which could be perceived as producing positive affect in anticipation of something pleasurable. Social martyr narratives, in contrast, are defined by the prospective martyr empathizing with those within the community who suffer, prompting a desire to alleviate the suffering of the community with an act of individual sacrifice. It emphasizes identity with the group and encourages people to place a group identity above their personal identity [45]. This definition connects with prior studies in which activity in dmPFC predicts persuasiveness in response to negative information (e.g., [27], [29]), with work associating ISCs in dmPFC with negative emotional valence [34], and with the general function of the mentalizing network in considering what other people are thinking and feeling (e.g., [21]).

Our results are the first to show a dissociation between how brain measures from reward/valuation regions and measures from mentalizing circuitry separately predict persuasion based on the narrative content of the stimuli. The finding that reward and mentalizing circuits can have dissociable effects on persuasion is consistent with some prior work (e.g., [29]). Similar results have also been observed in a content sharing task based on an instructional mindset manipulation [46]. Specifically, in that study, activity in the reward circuit only reliably increased relative to the control task when participants were instructed to choose content with a motivation to "Describe Yourself", which has a similar self-oriented focus as our Heroic narratives. Approaching the sharing task with the mindset to "Help Somebody" can be considered analogous to the community-oriented motivations evoked by our Social narratives. This framing was associated with increased activity in Self-focused and Mentalizing circuitry relative to the control task, with the "Describe Yourself" manipulation also increasing activity in

these regions. These results can be interpreted as being in alignment with the results of the present study.

There are some potential limitations to the external validity of our results. Participants were recruited from among University of Chicago students and the surrounding community, rather than among Muslims who are the real-life target audience for these videos. Furthermore, we were also not able to quantify the relative real-world effectiveness of these videos. ISIS does not share view counts publicly, and the videos are typically released in a decentralized manner across social media platforms and are often removed by content moderators and reposted from new accounts, making it very difficult to reliably estimate total engagement. It is clear, however, that these videos do play a role in recruiting fighters from Western countries based on surveys of ISIS recruits indicted for terrorism crimes in the United States [1]. Additionally, activity in the brain regions showing the strongest effects here, specifically VS/Nacc and dmPFC, have been associated with real-world engagement and persuasion in many prior studies (e.g., [16], [17], [18], [25], [26], [27]). Thus, it is plausible that these brain mechanisms drive persuasion in response to these videos in the real world.

Another complication is how to interpret the results in vmPFC. Both VS/NAcc and vmPFC have been identified as core regions in the brain's reward circuitry and value system [22], but vmPFC showed a modest effect across conditions driven by Social stimuli, more consistent with our findings in mentalizing regions such as dmPFC than with effects in VS/NAcc. Still, there are differences in the roles of VS/NAcc and vmPFC. For instance, in a study of crowdfunding campaigns, reward-related activity in both NAcc and mPFC predicted preferences within individuals, but only NAcc activity predicted out-of-sample campaign success [17]. This finding is consistent with the idea that mPFC integrates a wider range of inputs to decision-making, which could include idiosyncratic preference factors, while VS/NAcc activity reflects a more primary reward response [47]. Thus, fluctuation in vmPFC driven by reward might reflect more idiosyncratic factors than reward-driven fluctuations in VS/NAcc, and thus would not modulate ISCs. Additionally, vmPFC is associated with mentalizing as well as with reward (e.g., [21]), so the effects that are present in vmPFC may relate more to its role in mentalizing. It is therefore reasonable to interpret ISCs in VS/NAcc as reflecting more of a pure reward signal, while the role of vmPFC is more complex to interpret.

Ultimately, these results provide novel insight both on the specific topic of ISIS recruitment methods and more broadly on the neuroscience of persuasion. This work is one of the first to relate ISCs and ISFCs with persuasiveness of naturalistic stimuli. It also provides novel evidence in an area of interest to social neuroscientists in recent years: how reward/value and mentalizing networks work either independently or in concert to achieve persuasion. Finally, this work provides key support for the hypothesis that ISIS was able to recruit more broadly than other terrorist groups by employing two different routes for persuasion [4]. Persuasiveness of Heroic clips, which have been shown to appeal specifically to people without

strong ties to the Muslim community and individuals who are looking for personal glory, is associated with ISCs in a key node of the reward circuit. Persuasiveness of Social clips, which focus on threats to the Muslim community and appeal more to people with stronger ties to that community, is associated with ISCs in mentalizing and Default networks. To best counter ISIS propaganda, and to counter propaganda for other antisocial causes that might use a similar diversity of communicative appeals, will likely require distinct approaches for each type of narrative. The neural metrics used here could be relied upon to evaluate the effectiveness of counter-propaganda techniques in the future. Ultimately, this work is one piece of a broader effort to fight propaganda, and to make it more difficult for extremists to recruit people to participate in violent and antisocial behavior regardless of the political goal.

Methods

Participants

All participants signed informed consent forms prior to beginning the study. A total of 49 individuals completed the study. Of these, two were excluded due to having high levels of motion and low data quality on 3 out of 4 runs. Low data quality was defined, using output from MRIQC v22.0.6 [48], as having greater than 5% of volumes with framewise displacement (FD) > 0.9 mm and being at least 2 SD worse than the mean on tSNR, AFNI quality index, and mean FD. No other participants had more than one run meeting any of these criteria. An additional participant was excluded due to a screening failure: using a medication, Adderall, that was established as an exclusion criterion. Note that participants taking psychoactive medications were allowed in the study if they only reported taking a single SSRI anti-depressant. 46 participants thus contributed some data to the analyses. Of these, 34 individuals (mean age = 21.7 years, age range = 18-48 years, 14 male, 19 female, 1 non-binary) completed the full protocol. Three of these participants were missing post-scan questionnaire data for one video each; data from these videos were excluded from the primary data but included in the aggregate data. Data from the first 12 participants (mean age = 26.8 years, age range = 20-42, 4 male, 8 female), who were scanned while watching the videos but did not complete the post-scan questionnaire, were similarly included in aggregate data but excluded from other analyses. The first 10 participants also had a slightly different timing configuration, as noted below.

Behavioral Procedure

All procedures were approved as minimal risk by the University of Chicago Social and Behavioral Sciences (SBS) Institutional Review Board (IRB). Video clips were presented across four scan runs, with two videos (one Heroic and one Social) assigned to each run, in random order. Videos overlapped with those described in our previous work on this topic [5]. Specifically, the present study used videos labeled H01, H03, H06, and H07 for the Heroic condition and videos labeled S01, S02, S05, and S06 for the Social condition; see [5] for more

details. Heroic videos lasted 49 sec, 60 sec, 63 sec, and 61 sec, while Social videos lasted 34 sec, 55 sec, 38 sec, and 54 sec. Each run began with 4 seconds of fixation after the initial magnetic stabilization. Between the two videos, a fixation cross was presented lasting either 4 sec (first 10 subjects) or 16 sec (final 36 subjects). Additional rest time was also present at the end of each scan run, which differed based on the length of the specific videos in that run. The duration of each run was 126 s (first 10 subjects) or 148 s (final 36 subjects).

As part of the pre-scan screening procedure, participants completed an 8-item Justice Sensitivity measure [38] and the 5-item DUREL religiosity index [39]. In the scanner, the video clip task described presently was preceded by an unrelated moral reasoning task similar to [49]; fMRI data from that task will be reported elsewhere. As part of the pre-scan screening procedure, only prospective participants who showed a sufficient range of support and moral conviction ratings on the tested sociopolitical issues for the moral reasoning task were recruited into the study. After the scan, participants were given the opportunity to watch each video again. Immediately after watching each video, they were asked to answer 4 questions on a 7-point Likert scale: "This video would help the militant group recruit" (Persuasiveness), "I could picture myself in the scene of the events described in the video." (Narrative transportation 1), "After the video ended, I found it easy to put it out of my mind." (Narrative transportation 2, reverse-scored), and "I found my mind wandering while watching the video." (Narrative transportation 3, reverse-scored). Scores on the final 3 questions were averaged to generate a measure of narrative transportation for each video.

Neuroimaging Procedure

Functional MRI data were collected on a Philips Achieva 3 T MRI scanner at the University of Chicago MRI Research Center. Functional scans used a 2000 ms TR length, 28 ms TE length, flip angle 80°, 40 ascending slices with a 0.3 mm gap between slices. Voxel size was 3.0 mm x 3.1 mm x 3.0 mm, with a 64 x 62 matrix, and field of view of 192 x 192 mm^2. Scans for the first 10 participants had 63 volumes, while scans for the final 36 participants had 74 volumes. A T1-weighted structural image was also collected, with TR length = 8 ms, TE length = 3.5 ms, flip angle 8°, 0.85 mm x 0.85 mm x 0.85 mm voxels, and a 284 x 260 matrix. For field inhomogeneity mapping, two short runs were collected using the same parameters as the functional runs, with 5 volumes collected in the anterior → posterior direction and 5 volumes collected in the posterior → anterior direction.

fMRI Preprocessing

Data were preprocessed using fmriprep v22.1.1 [50]. The following 4 paragraphs are excerpted and adapted from the documentation distributed with fmriprep. The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection [51], distributed with ANTs 2.3.3 [52], and used as T1w-reference throughout the workflow. The

T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (part of FSL 6.0.5.1) [53]. Brain surfaces were reconstructed using recon-all (FreeSurfer 7.2.0) [54], and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle [55]. Volume-based spatial normalization to standard space was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following template was used for spatial normalization: *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model* (TemplateFlow ID: MNI152NLin6Asym) [56]. A B_0-nonuniformity map (or *fieldmap*) was estimated based on two echo-planar imaging (EPI) references with topup (FSL 6.0.5.1) [57].

For each BOLD run, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using mcflirt (FSL 6.0.5.1) [58]. The estimated *fieldmap* was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run. The field coefficients were mapped on to the reference EPI using the transform. BOLD runs were slice-time corrected to 0.975s (0.5 of slice acquisition range 0s-1.95s) using 3dTshift from AFNI [59]. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration [60]. Co-registration was configured with six degrees of freedom.

A set of physiological regressors were extracted to allow for component-based noise correction (a*CompCor*) ([61], [62]). Principal components were estimated after high-pass filtering of the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off). Probabilistic masks for CSF and WM were generated in anatomical space, and components were calculated separately within each mask.

The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin6Asym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. All resamplings can be performed with *a single interpolation step* by composing all pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels [63]. Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

fMRI Data Analysis

A selection of the confound regressors obtained from fmriprep were applied to the preprocessed data using the Nilearn clean_img function. The 6 motion parameter time series derived from head motion estimates calculated in the correction step were expanded with the inclusion of temporal derivatives and quadratic terms [64], yielding a total of 24 motion parameters (6 basic parameters plus squared, derivatives, and squared derivatives of each one). The confound model also included the first 5 aCompCor parameters from WM, and up to the first 5 aCompCor parameters from CSF, though sometimes as few as 3 CSF parameters were used if fewer than 5 components were needed to account for 50% of variance in CSF signal. A cosine basis regressor was included for high-pass filtering if available. For the first 10 participants, a high pass filter of .008 Hz was applied in clean_img because the run was too short to include cosine basis functions. The first volume from each run was dropped in order to allow for use of derivatives of motion parameters as regressors.

Data from each run was resampled to match the ROI space using the Nilearn resample_to_img function. The time series for each video was then extracted from the cleaned images using the Nibabel slicer function, beginning 2 volumes after stimulus onset and ending 2 volumes after stimulus offset, to account for 4 s of delay in the hemodynamic response. Each ROI was applied as a mask using the Nilearn NiftiMasker function, and BOLD signal was averaged for all voxels at a given time point within the mask to yield a 1-dimensional time series for that individual and video. Each ROI was cropped using the whole-brain mask generated by fmriprep for the relevant scan run. If fewer than 10% of the voxels within an ROI were included in the whole-brain mask for a given participant and run, no data was returned and that individual was excluded from all analyses of the ROI for that video.

Intersubject correlations (ISCs) were computed as the Pearson correlation coefficient between an individual's time series for a specific ROI and video clip and the mean time series for all 45 other individuals (or however many individuals had valid data in that ROI, if not the full sample) for the same ROI and video clip. An arctanh (Fisher's Z) transformation was then applied to the resulting correlation coefficient. Intersubject functional connectivity (ISFC) for each pair of regions was computed in a similar manner, except that an individual's time series was compared to the mean time series across all individuals collected from the same clip but from a *different* ROI.

ROIs were defined based on prior literature. For reward regions, the VS/NAcc and vmPFC ROIs were defined from a meta-analysis of domain-general reward response, as shown in Figure 9 of [22], obtained from www.kablelab.com . Each cluster was further separated into R and L hemisphere segments by a sagittal split at the midline. A VTA ROI was defined based on a probabilistic atlas [40] as all voxels with at least a 10% chance of being in the VTA. This ROI was also split by hemisphere via a sagittal split at the midline. For all reward regions, ISCs were initially computed separately within each hemisphere and then averaged. To identify brain

regions associated with mentalizing, spheres with radius 10 mm were constructed using *fslmaths* around reported peak coordinates from published studies. Specifically, spheres were created around all 16 peaks in Table 1 of the meta-analysis reported by [21]. A total of 14 ROIs were generated, as two pairs of spheres (in R TPJ and L precentral gyrus) overlapped, and thus these paired spheres were combined into single ROIs. For the alternative meta-analysis [41], the peak coordinate for each of 9 clusters reported in Table 3 from studies on adults were used to create ROIs. Finally, the third definition of mentalizing was created from the 13 peaks (excluding cerebellum) from the Why > How contrast in Table 2 from [42]. Here, the R dmPFC sphere overlapped slightly with one of the L dmPFC spheres, so the overlapping ROIs generated from these two peaks were separated at the sagittal midline. Additionally, two other peaks in R anterior temporal cortex overlapped and were combined into a single ROI, yielding 12 total ROIs.

Acknowledgements

Funding was provided by a United States Department of Defense Minerva Initiative grant (FS9550-16-1-0074) titled "The Social and Neurological Construction of Martyrdom" to R.P. and J.D. We thank Monica Rosenberg and Hayoung Song for guidance on naturalistic data analysis methods. Keith Yoder, Qiongwen (Jovie) Cao, Tongyu Gu, Allie Reynolds, and Emma Van Steertegem contributed to participant recruitment, study implementation, and data collection for this project.

Competing interests

The authors declare no competing interests.

Author Contributions

J.D., R.P., K.R. – designed research; M.C. – collected data; M.C. and Y.C.L. – data analysis; M.C. – initial manuscript draft; M.C., Y.C.L., K.R., R.P., J.D. – manuscript revisions

Data Availability Statement

Behavioral data, neuroimaging data extracted from regions of interest, and R code for statistical analyses are available at https://osf.io/w5qpx . Other data and code are available upon request from the corresponding authors.

References

1. Pape, R. A., Decety, J., Ruby, K., Rivas, A. A., Jessen, J., & Wegner, C. (2017). "The American Face of ISIS: Analysis of ISIS-Related Terrorism in the US March 2014-August 2016." Australian Strategic Policy Institute. Link: https://www.aspi.org.au/report/american-face-isis-analysis-isis-related-terrorism-us-march-2014-august-2016
2. Dodwell, B., Milton, D., & Rassler, D. (2016). "Then and now: Comparing the flow of foreign fighters to AQI and the Islamic State." West Point, NY: Combatting Terrorism Center. Link: https://ctc.westpoint.edu/wp-content/uploads/2016/12/Then-and-Now.pdf
3. Pape, R. A., Rovang, D., Ruby, K. G., and Decety, J. (2018). "Mobilizing to Martyrdom: A Narrative Theory of High-Risk Mobilization." *APSA Preprints.* Available at https://doi.org/10.33774/apsa-2023-x660r .
4. Pape, R. A., Decety, J., Ruby, K. G., Yoder, K., & Rovang, D. (2023). "Identity and narrative persuasion: How ISIS Western-directed propaganda works." *APSA Preprints.* Available at https://doi.org/10.33774/apsa-2023-68cbq .
5. Yoder, K., Ruby, K., Pape, R., & Decety, J. (2020). EEG distinguishes heroic narratives in ISIS online video propaganda. *Scientific Reports, 10,* 19593.
6. Frischlich, L., Rieger, D., Morten, A., & Bente, G. (2018). The power of a good story: Narrative persuasion in extremist propaganda and videos against violent extremism. *International Journal of Conflict and Violence (IJCV),* 12, a644.
7. Baumert, A., Rothmund, T., Thomas, N., Gollwitzer, M., & Schmitt, M. (2013). Justice as a moral motive: Belief in a just world and justice sensitivity as potential indicators of the justice motive. In Heinrichs, K., Oser, F. K., & Lovat, T. (eds.), *Handbook of moral motivation* (pp. 159-179). Brill.
8. Yoder, K. J., & Decety, J. (2014a). The good, the bad, and the just: Justice sensitivity predicts neural response during moral evaluation of actions performed by others. *Journal of Neuroscience, 34,* 4161-4166.
9. Yoder, K. J., & Decety, J. (2014b). Spatiotemporal neural dynamics of moral judgment: a high-density ERP study. *Neuropsychologia, 60,* 39-45.
10. Decety, J., & Yoder, K. J. (2017). The emerging social neuroscience of justice motivation. *Trends in Cognitive Sciences, 21,* 6-14.
11. Decety, J., & Yoder, K. J. (2016). Empathy and motivation for justice: Cognitive empathy and concern, but not emotional empathy, predict sensitivity to injustice for others. *Social Neuroscience, 11,* 1-14.
12. Rothmund, T., Bromme, L., & Azevedo, F. (2020). Justice for the people? How Justice Sensitivity can foster and impair support for populist radical-right parties and politicians in the United States and in Germany. *Political Psychology, 41,* 479-497
13. Cloutier, J., Heatherton, T. F., Whalen, P. J., & Kelley, W. M. (2008). Are attractive people rewarding? Sex differences in the neural substrates of facial attractiveness. *Journal of Cognitive Neuroscience, 20,* 941-951.
14. Levy, I., Lazzaro, S. C., Rutledge, R. B., & Glimcher, P. W. (2011). Choice from non-choice: Predicting consumer preferences from Blood Oxygenation Level-Dependent signals obtained during passive viewing. *Journal of Neuroscience, 31,* 118-125.

15. Tusche, A., Kahnt, T., Wisniewski, D., & Haynes, J.-D. (2013). Automatic processing of political preferences in the human brain. *NeuroImage, 72,* 174-182.
16. Scholz, C., Baek, E. C., O'Donnell, M. B., Kim, H. S., Cappella, J. N., & Falk, E. B. (2017). A neural model of valuation and information virality. *Proceedings of the National Academy of Sciences, 114,* 2881–2886.
17. Genevsky, A., Yoon, C., Knutson, B. (2017). When brain beats behavior: Neuroforecasting crowdfunding outcomes. *Journal of Neuroscience, 37,* 8625-8634.
18. Tong, L.C., Acikalin, M.Y., Genevsky, A., Shiv, B., Knutson, B. (2020). Brain activity forecasts video engagement in an internet attention market. *Proceedings of the National Academy of Sciences, 117,* 6936-6941.
19. Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current Biology, 20,* 1165-1170.
20. Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological Science, 22,* 894-900.
21. Arioli, M., Cattaneo, Z., Ricciardi, E., & Canessa, N. (2021). Overlapping and specific neural correlates for empathizing, affective mentalizing, and cognitive mentalizing: A coordinate-based meta-analytic study. *Human Brain Mapping, 42,* 4777-4804.
22. 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.
23. Cacioppo, J. T., Cacioppo, S., & Petty, R. E. (2018). The neuroscience of persuasion: A review with an emphasis on issues and opportunities. *Social Neuroscience, 13,* 129-172.
24. Falk, E. B., Rameson, L., Berkman, E. T., Liao, B., Kang, Y., Inagaki, T., & Lieberman, M. D. (2010). The neural correlates of persuasion: A common network across cultures and media. *Journal of Cognitive Neuroscience, 22,* 2447-2459.
25. Chua, H. F., Ho, S. S., Jasinska, A. J., Polk, T. A., Welsh, R. C., Liberzon, I., & Strecher, V. J. (2011). Self-related neural response to tailored smoking-cessation messages predicts quitting. *Nature Neuroscience, 14,* 426–427.
26. Wang, A. L., Ruparel, K., Loughead, J. W., Strasser, A. A., Blady, S. J., Lynch, K. G., Romer, D., Cappella, J. N., Lerman, C., & Langleben, D. D. (2013). Content matters: Neuroimaging investigation of brain and behavioral impact of televised anti-tobacco public service announcements. *Journal of Neuroscience, 33,* 7420–7427.
27. Falk, E. B., O'Donnell, M. B., Tompson, S., Gonzalez, R., Dal Cin, S., Strecher, V., Cummings, K.M., & An, L. (2016). Functional brain imaging predicts public health campaign success. *Social Cognitive and Affective Neuroscience, 11,* 204-214.
28. Welborn, B. L., Lieberman, M. D., Goldenberg, D., Fuligni, A. J., Galván, A., & Telzer, E. H. (2015). Neural mechanisms of social influence in adolescence. *Social Cognitive and Affective Neuroscience, 11,* 100-109.
29. Baek, E. C., O'Donnell, M. B., Scholz, C. Pei, R., Garcia, J. O., Vettel, J. M., & Falk, E. B. (2021). Activity in the brain's valuation and mentalizing networks is associated with propagation of online recommendation. *Scientific Reports, 11,* 11196.
30. Hasson, U., Malach, R., & Heeger, D. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences, 14,* 40–48.

31. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303,* 1634-1640.
32. Nastase, S., Gazzola, V., Hasson, U., & Keyers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience, 14,* 667–685.
33. Simony, E., Honey, C., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications, 7,* 12141.
34. Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., & Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences, 109,* 9599-9604.
35. Imhof, M. A., Schmälzle, R., Renner, B., & Schupp, H. T. (2017). How real-life health messages engage our brains: Shared processing of effective anti-alcohol videos. *Social, Cognitive, and Affective Neuroscience, 12,* 1188-1196.
36. Schmälzle, R., Hacker, F. E. K., Honey, C., & Hasson, U. (2015). Engaged listeners: Shared neural processing of powerful political speeches. *Social Cognitive and Affective Neuroscience, 10,* 1137–1143.
37. Grall, C., Weber, R., Tamborini, R., & Schmälzle, R. (2021). Stories collectively engage listeners' brains: Enhanced intersubject correlations during reception of personal narratives. *Journal of Communication, 71,* 332-355.
38. Baumert, A., Beierlein, C., Schmitt, M., Kemper, C. J., Kovaleva, A., Liebig, S., & Rammstedt, B. (2014). Measuring four perspectives of justice sensitivity with two items each. *Journal of Personality Assessment, 96,* 380-390.
39. Koenig, H. G., Meador, K. G., & Parkerson, G. (1997). Religion index for psychiatric research. *American Journal of Psychiatry, 154,* 885–886.
40. Murty, V. P., Shermohammed, M., Smith, D. V., Carter, R. M., Huettel, S. A., Adcock, R. A. (2014). Resting state networks distinguish human ventral tegmental area from substantia nigra. *Neuroimage, 100,* 580-589
41. Fehlbaum, L. V., Borbás, R., Paul, K., Eickhoff, S. B., & Raschle, N. (2022). Early and late neural correlates of mentalizing: ALE meta-analyses in adults, children and adolescents. *Social Cognitive and Affective Neuroscience, 17,* 351-366.
42. Spunt, R. P., & Adolphs, R. (2014). Validating the Why/How contrast for functional MRI studies of Theory of Mind. *NeuroImage, 99,* 301-311.
43. Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex, 28,* 3095-3114.
44. Zalesky, A., Fornito, A., & Bullmore, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage, 53,* 1197-1207.
45. Pape, R. A. (2005). *Dying to win: The strategic logic of suicide terrorism.* New York: Random House.
46. Scholz, C., Baek, E., & Falk, E. B. (2023). Invoking self-related and social thoughts impacts online information sharing. *Social Cognitive and Affective Neuroscience, 18*, nsad013.
47. Samanez-Larkin, G., & Knutson, B. (2015). Decision making in the ageing brain: Changes in affective and motivational circuits. *Nature Reviews Neuroscience, 16,* 278-289.

48. Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., Gorgolewski, K. J. (2017) MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE, 12,* e0184661.
49. Yoder, K. J., & Decety, J. (2022). Moral conviction and metacognitive ability shape multiple stages of information processing during social decision-making. *Cortex, 151,* 162-175.
50. Esteban, O., Ciric, R., Finc, K. et al. (2020). Analysis of task-based functional MRI data preprocessed with fMRIPrep*. Nature Protocols, 15,* 2186–2202.
51. Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging, 29,* 1310-1320.
52. Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis, 12,* 26–41.
53. Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm." *IEEE Transactions on Medical Imaging, 20,* 45–57.
54. Dale, A. M., Fischl, B., Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage, 9,* 179–194.
55. Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., et al. (2017). Mindboggling Morphometry of Human Brains. *PLOS Computational Biology, 13,* e1005350.
56. Evans, A. C. , Janke, A. L., Collins, D. L., & Baillet, S. (2012). Brain templates and atlases. *NeuroImage, 62,* 911–922.
57. Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *NeuroImage, 20,* 870–888.
58. Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images." *NeuroImage 17,* 825–841.
59. Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine, 10,* 171–178.
60. Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage, 48,* 63–72.
61. Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI." *NeuroImage, 37,* 90–101.
62. Muschelli, J., Nebel, M. B., Caffo, B. S., Barber, A. D., Pekar, J. J., & Mostofsky, S. H. (2014). Reduction of motion-related artifacts in resting state fMRI using aCompCor. *NeuroImage, 96,* 22-35.
63. Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis, 1,* 76–85.
64. Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., et al. (2013). An improved framework for confound regression and

filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage, 64,* 240–256.