Identifying A Neural Signature That Predicts Self-Focus

Danika Geisler and Meghan L. Meyer
Columbia University

Corresponding Author:
Meghan L. Meyer
meghan.meyer@columbia.edu
Department of Psychology
Columbia University

Abstract

People are remarkably self-focused, disproportionately choosing to think about themselves relative to other topics. Self-focus can be adaptive, helping individuals fulfill their needs. It can also go haywire, with maladaptive self-focus a risk and maintenance factor for internalizing disorders like depression. Yet, the neural mechanism driving people to focus on themselves remains unknown. This gap is due to timing: while prior research measures neural activity the moment participants are instructed to self-reflect, a brain state that precedes, or nudges, the bias to spontaneously focus on the self remains undetermined. We identified a default network neural signature from pre-trial activity that predicts 1) multiple indicators of self-focus within our sample and 2) internalizing symptoms in a separate sample from the Human Connectome Project. This is the first work to “decode” the bias to focus on the self and paves the way towards stopping maladaptive self-focus in its course.

Writer David Foster Wallace once referred to the self as "our default setting"[1]. Psychological data support his view. Across cultures, people disproportionately think about themselves while mind wandering[2–4]. They also are more likely to remember and communicate self-relevant information than information unrelated to the self[5–11]. Even when people try to take the perspective of someone dissimilar to themselves, much of the time they still end up projecting their own point of view[12,13]. While self-focus is necessary and positive in some forms—such as detecting our hunger cues or social needs—in its most pernicious forms, self-focus is a risk and maintenance factor for internalizing disorders such as depression[14–19]. Identifying what drives the bias towards focusing on ourselves would pave the way towards stopping maladaptive self-focus in its course, in turn preventing its negative downstream consequences.

Yet, the underlying neural mechanism(s) generating the bias towards self-focus remains to be determined. This gap is due, in large part, to timing. Extensive neuroscience research on "the self" captures neural activity *while* participants are instructed to think about themselves[20], robustly implicating the medial prefrontal cortex, Brodmann Area 10 (MPFC/BA10) in active self-reflection. This area of research has been highly generative but fails to identify neural processes that set self-focus in motion. Meeting that goal would require examining neural activity *before* people spontaneously focus on themselves and assessing whether such neural patterns temporally predict self-focus. This alternative approach fits with predictive coding accounts of brain function, which broadly suggest endogenous, default brain states predict subsequent perception and cognition[21–23]. For example, pre-stimulus fusiform gyrus activity predicts which of two competing visual stimuli is perceived[24] and pre-stimulus hippocampal patterns shape stimulus encoding and memory[25–27].

The possibility that pre-stimulus neural processes predict the bias towards self-focus is further bolstered by the fact that MPFC/BA10 is a key node of the default network, known to robustly engage "by default", without the presence of any stimuli[28]. Specifically, the default network shows stronger neural activity while passively resting relative to many experimental, cognitive tasks. Indeed, many scholars have speculated that engaging MPFC/BA10, and the default network more generally, during stimulus-free rest may reflect some form of self-focus in the scanner[28,29]. That said, while past work has shown greater resting state functional connectivity (i.e., time-course correlations between brain regions)

between MPFC/BA10 and other default network regions correlates with self-focus[30,31], the direction of this relationship is unknown. Moreover, previous work focuses on functional connectivity to investigate neural activity during rest. Yet functional connectivity is too coarse a metric to parse whether a neural signal at rest captures active self-reflection versus triggers the bias towards self-focus.

We took a multipronged approach to determine the neural signature predicting the bias to focus on ourselves. First, we developed a paradigm designed to behaviorally measure the bias towards self-focus. In this paradigm, participants believed they were deciding which kinds of experimental trials they would get in a separate task that would immediately follow. The structure of the decision task was as follows: it started with a brief (~4.5 secs) pre-trial rest period followed by a trial in which participants choose if they would later like to think about themselves, a close other (i.e., nominated friend), or a well-known other (i.e., President Biden) across multiple dimensions assessed separately (i.e., personality and physical traits; social roles; preferences; past and future). Each choice was followed by an attention reorienting trial (i.e., shape matching) to help ensure participants cleared their mind before the next pre-trial rest period (see Figure 1A). Participants completed this task while undergoing functional magnetic resonance imaging (fMRI) and their behavioral responses indeed reflected a bias towards self-focus: they disproportionately choose to think about themselves (vs. the close and well-known others) and were preferentially faster to do so (see Results).

Critical to the question of what may predict self-focus, we assessed neural responses to each pre-trial rest phase of this task and examined how it modulated the speed and decision to choose the self (vs. others) on the immediately following trial. To date, only one study has probed pre-stimulus neural activity in a self-reflection task, finding that, on a trial-by-trial basis, faster responses to questions assessing beliefs about one's traits (e.g., "Am I funny?" yes/no) are preceded by stronger MPFC/BA10 activity[32]. This finding is consistent with the present paper's hypothesis: it shows that "defaulting" to MPFC/BA10 facilitates access to self-views. However, necessary and sufficient support for the bias towards self-focus requires that default MPFC/BA10 neural responses predict the preference to focus on the self (vs. others). To this end, we trained support vector machine (SVM) classifiers in MPFC/BA10, and the default network more generally, to determine if multivariate patterns during pre-trial rest predict the following choice to think about the self

(vs. others). This approach allowed us to determine neural patterns that "decode" the desire to think about the self before the decision has been made. We call this neural signature the "pre-self pattern".

We next investigated if our pre-self pattern generalizes to predict self-focus in another context—an extended resting state scan. We assessed this possibility in two ways. First, in the very beginning of our experiment, participants completed an 8-minute resting state scan and rated the extent to which they were focused on themselves, others, the past, and the future every 2 minutes of the rest period. We took the pre-self pattern identified in the pre-trial rest period from the task described above and applied it to the resting state data. Specifically, we used multivariate pattern similarity analysis to quantify the extent to which the pre-self pattern was present during each second of the two minutes preceding participants' self-reports during the resting state scan. The approach taken here is similar to "reinstatement" analysis from the memory consolidation literature, in which researchers assess the extent to which multivariate patterns from encoding reappear during post-encoding rest[33,34]. Here, we refer to the approach as "instatement" analysis because we are investigating if a pre-self pattern (rather than a pattern for something previously encoded) spontaneously appears during a baseline rest scan. We reasoned that if we discovered a neural signature that predicts self-focus, then the strength of its presence during extended rest should preferentially predict self-reported self-focus, too.

Second, a neural signature that predicts self-focus should also temporally predict the neural signature that captures active self-reflection. That is, if the pre-self pattern nudges self-reflection, then it should nudge the neural pattern reflecting active self-reflection. To examine this, we capitalized on a separate task completed by our participants. At the end of this experiment, participants completed a traditional self-reflection task, in which they rated their own personality traits. This approach has been used numerous times in the social neuroscience literature to capture active self-reflection[35–40]. For each participant, we derived their multivariate pattern (in the MPFC/BA10 and default network more generally) associated with active, instructed self-reflection. We then went back to our resting state data and examined whether the strength of the pre-self pattern temporally predicted the presence of the active self-reflection pattern (and not vice versa). Such a result would be consistent with the hypothesis that the pre-self pattern predicts active self-reflection. More broadly, positive results from our multi-pronged approach would provide robust evidence

158 that we identified a neural signature that predicts multiple characterizations of self-focus
159 (i.e., decisions, self-report, neural).
160
161 Moving back to the larger clinical implications of this work, self-focused thought is a risk
162 and maintenance factor for internalizing disorders, including depression and anxiety[14–19].
163 If we are able to identify a neural signature that allows us to better understand what nudges
164 individuals into these maladaptive states, it would pave the way towards stopping
165 maladaptive self-focus in its course, in turn preventing its negative downstream
166 consequences. Additionally, if we have truly identified a neural signature, we should be
167 able to apply it to individuals from separate datasets and see that it relates to participants'
168 self-focus. We thus tested whether our pre-self pattern predicts outcomes related to
169 maladaptive self-focus in data from the Human Connectome Project. The Human
170 Connectome Project dataset includes a resting state scan and internalizing score for each
171 participant, which reflects symptoms like anxiety and depression that are related to self-
172 focus. An intersubject representational similarity analysis (IS-RSA[41]) revealed that
173 individuals with higher internalizing scores rhythmically default to the pre-self pattern. In
174 contrast, individuals low in internalizing do not show any systematic structure to the
175 presence of their pre-self pattern; they move into and out of the pre-self pattern
176 idiosyncratically. These results not only show the pre-self pattern can meaningfully predict
177 maladaptive forms of self-focus in another sample. They further identify a novel way to
178 think about how internalizing relates to brain function. Internalizing may be a maladaptive
179 form of self-focus, at least in part, because individuals high on this dimension consistently
180 return to the pre-self pattern while mind wandering, making self-focus inescapable.

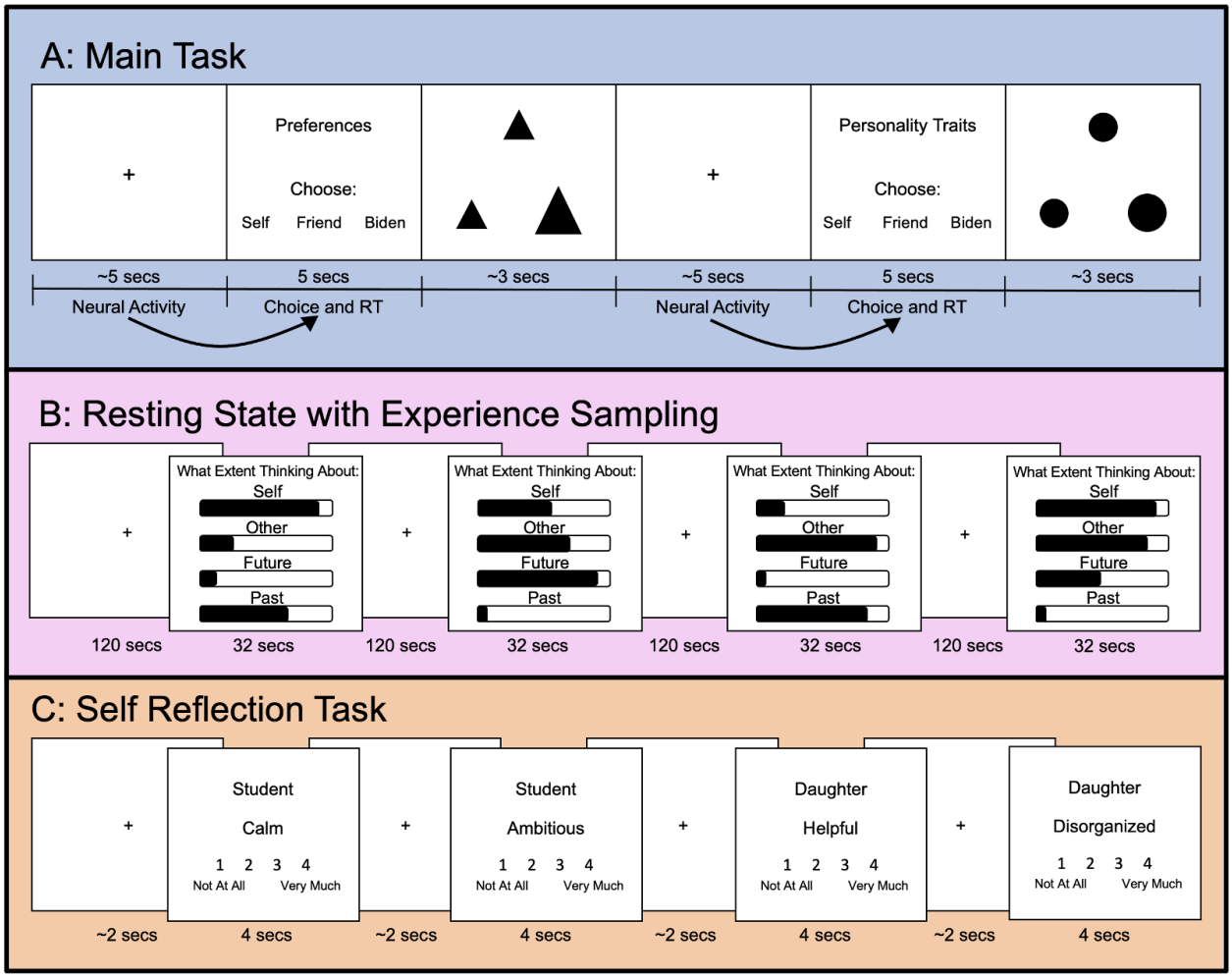

Fig. 1 A) Our main task design included 3 parts that repeated. First was a pre-trial jittered rest period (2.5–6 sec, mean = 4.5 sec); second was the choice activity where participants choose who (themselves, a designated friend, or Biden) they want to think about in a later task (5 sec); third, a shape matching task, aimed at moving participants' minds off their last choice, followed. These three parts repeated 54 times in run 1 and 54 times in run 2 for a total of 108 trials. Our analysis examined neural activity during the pre-trial jittered rest in relation to the subsequent task choice and response time. B) The first fMRI run that the participants completed was an 8 minute long resting state scan. The 8 minutes were broken up into four, 2 minute long sections. After each 2 minute section, participants had 32 seconds to rate the extent to which they were thinking about themselves, others, the future, and the past. These ratings were made on a continuous scale with ‘not at all’ on one end, ‘completely’ on the other end, and ‘somewhat’ in the middle of the scale. C) The final task participants completed was a typical self reflection task. In this task participants

were instructed to rate how well an adjective describes their personality in different social roles (Friend, Student, Significant Other, Son/Daughter, and Worker) on a scale from 1 to 4 using button boxes. We used the Big 5 list of 100 adjectives[42]. The rating trials were 4 seconds long and the jittered rest was 1–3 sec, mean = 2 sec. There were two runs of 101 trials each, with a total of 202 trials.

Results

When given the choice, people prefer to think about themselves over others

In our primary fMRI task, participants believed they were choosing trials that would appear in their subsequent fMRI task. Specifically, they choose if they would later like to think about themselves, a close other (i.e., nominated friend), or a well-known other (i.e., President Biden) across multiple dimensions assessed separately (i.e., personality and physical traits; social roles; preferences; past and future), with a total of 108 choices made (See Figure 1A). Participants' decisions indicated a strong predisposition to default towards self-focus. A repeated-measures ANOVA demonstrated a main effect of choice on the number of trials selected ($F(2, 62) = 29.26$, $p < .001$). Follow-up, paired sample *t*-tests showed participants choose to think about themselves significantly more than a designated friend ($t(31) = 3.76$, $p < 0.001$, Cohen's d = .67) and Biden ($t(31) = 7.09$, $p < 0.001$, Cohen's d = 1.25). There was also a main effect of choice on response time ($F(2, 3431.3) = 11.97$, $p < .001$) such that participants were faster to make decisions to think about themselves in comparison to a friend ($\beta = 0.11$, standardized $\beta = 0.02$, $t(3429) = 4.68$, $p = .003$) and Biden ($\beta = 0.08$, standardized $\beta = 0.03$, $t(3435) = 2.93$, $p < 0.001$). Notably, participants were also more likely to choose to think about their friend than Biden ($t(31) = 4.37$, $p < 0.001$, Cohen's d = .77), and were faster to choose their friend over Biden ($\beta = 0.08$, standardized $\beta = 0.03$, $t(3435) = 2.93$, and $p < 0.001$).

Two follow-up analyses further assessed the pervasiveness of the bias towards self-focus. One possibility is that self-focus is driven by a particular dimension assessed, rather than a generalizable preference to think about oneself. For example, perhaps the desire to think about the self is driven specifically by participants' preference to think about their past. If so, then we would have identified a nuanced bias towards self-focus rather than a more general one. To assess this possibility, we tested whether choice preferences regarding who participants wanted to think about (self/friend/Biden) interacted with the content of the dimension considered (social roles, preferences, physical traits, personality traits, future,

and past). Choice preferences, but not reaction time, significantly interacted with dimensions (choice: $F(10, 310) = 3.00$, $p = .001$, $\eta_p^2 = 0.088$; response time: $F(10, 3404.2) = 1.75$, $p = .064$, $\eta_p^2 = 0.002$). Critically, subsequent t-tests revealed that there was not a dimension within self choices that was significantly different than any of the others (*t*'s < 2.7; *p*'s > 0.50, bonferonni-corrected). Instead, the interaction was driven by choices for the *other targets*. When participants chose to later think about their friend, they were more likely to choose to think about their friend's personality traits than social roles ($t(31) = 4.20$, $p = 0.009$, Cohen's d = 0.87). When participants chose to think about Biden, they were more likely to choose to think about his social roles than preferences and past ($t_{preferences}(31) = 4.00$, $p = 0.017$, Cohen's d = 0.87; $t_{past}(31) = 3.50$, $p = 0.045$, Cohen's d = 0.82), and more likely to choose to think about his future than preferences ($t(31) = 3.88$, $p = 0.023$, Cohen's d = 0.66). In other words, self-focus was not driven by a particular dimension(s), although when people preferred to think about their friend or Biden, certain dimensions were more preferred than others.

Another possibility is that the bias towards self-focus only occurs initially, in the beginning of the task, but this preference dissipates over time. This would suggest that people are not overwhelmingly self-focused, rather they can correct for this bias. However, there was no significant linear trend of choosing self over the length of the entire, two-run task ($\beta = 0.04$, standardized $\beta = 0.11$, $t(3447) = 0.40$, $p = 0.69$) nor for run one vs. run two ($\beta = -0.05$, standardized $\beta = 0.07$, $t(3447) = -0.67$, $p = 0.50$). Similarly, there was no significant choice by run interaction ($\beta = -0.01$, standardized $\beta = 0.07$, $t(3447) = -0.16$, $p = .87$). The same analyses for the other targets (i.e., friend; Biden) indicated that choice behavior also did not change with time for decisions to think about the friend (β's < .06; *p*'s > .60) or Biden (β's < .14; *p*'s > .14). Collectively, the behavioral results suggest a strong, persistent predisposition to quickly choose to focus on the self (vs. others).

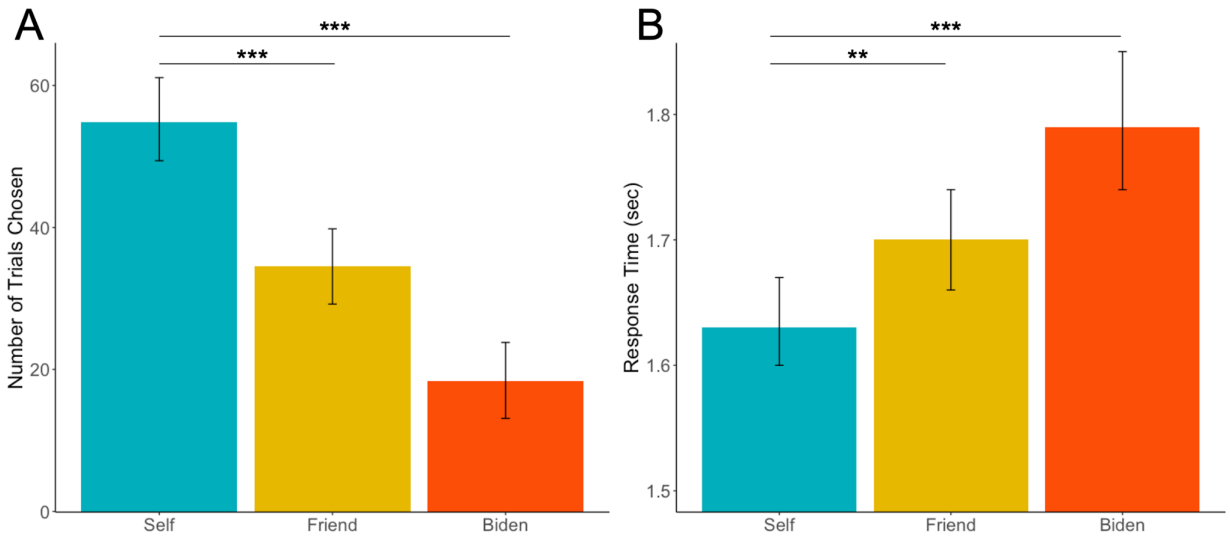

Fig. 2 (A) Participants choose to think about themselves significantly more than their self-nominated friend ($t(31) = 3.76$, $p < 0.001$, Cohen's d = .67) or Biden ($t(31) = 7.09$, $p < 0.001$, Cohen's d = 1.25). (B) Participants are faster in their decisions to think about themselves in comparison to a friend (β = 0.11, standardized β = 0.02, $t(3429) = 4.68$, $p = .003$) and Biden (β = 0.08, standardized β = 0.03, $t(3435) = 2.93$, $p < 0.001$). ***indicates $p < .001$; **indicates $p < .005$.

A neural signature that predicts self-focus: Evidence from the brief, pre-trial rest in the choice task

Mean MPFC/BA10 Activity during Pre-Trial Rest Parametrically Modulates Decisions to Think about the Self

The primary goal of the choice task was to derive a neural signature that precedes and decodes self-focus. However, we first wanted to assess if we conceptually replicate the single, previous study investigating the role of pre-stimulus responses on self-processing[32]. Consistent with that prior work, we performed a parametric modulation analysis in which neural activity during pre-trial rest was modulated—on a trial-by-trial basis—by the speed and choice of the next trial. Note that these and all subsequent analyses are run on the residual images from the task activation models to ensure pre-trial rest activity is not contaminated by task-evoked effects. Given that the previous work found the magnitude of MPFC/BA10 activity preferentially facilitates the speed with which

participants answer questions about themselves, we probed the parametric modulation analysis here in an MPFC/BA10 region-of-interest (ROI) predefined by Yeo et al.[43]; see Methods). Consistent with the prior work, faster decisions to choose the self (vs. friend and Biden) corresponded with greater mean activity in the MPFC/BA10 during the previous pre-trial rest period ($t(31) = -2.20$, $p = 0.036$, Cohen's d = -0.39). No additional clusters emerged in follow-up whole-brain parametric modulation analyses searching for brain regions whose greater activation during pre-trial rest predicted faster decisions to think about: 1) self vs other (friend and Biden), 2) friend vs. self, 3) Biden vs. self or 4) friend vs. Biden). Conceptually replicating prior work, the results indicate that faster choices to think about the self are preceded by greater mean MPFC/BA10 activity.

Multivariate Neural Patterns in the Core Default Network Subsystem during Pre-Trial Rest Predict Decisions to Think about the Self

Next, we returned to our primary goal – to test whether neural patterns during pre-trial rest could "decode" if participants next wanted to think about the self. In the decoding analysis presented here, we used the decision on the choice task as the outcome variable to determine if neural patterns during pre-trial rest predict the subsequent choice to think about the self. Specifically, using multi-voxel pattern analysis (MVPA) on the pre-trial rest, we trained a linear support vector machine (SVM) classifier to differentiate between subsequent self-choices vs. other-choices (friend and Biden). Friend and Biden choices were combined so there was a roughly equal number of self vs. other trial types, which helps ensure the analysis is unbiased. We first performed the analysis and validation specifically within the Yeo et al.[43] MPFC/BA10 ROI. We computed prediction performance using the 6-fold balanced cross-validation procedure[44,45]. Specifically, we subdivided the data into 6 separate folds (5-6 participants in each group) and used all of the data except for one-fold to train the model and then tested the model using the left-out fold. We then iterated over this process for each fold and an average classification accuracy was calculated. A null distribution was computed using 10,000 permutations of 6-fold randomized SVMs and p-values were calculated to indicate statistical significance of the predictive accuracy (see Methods).

The approach identified that distributed, pre-stimulus patterns in MPFC/BA10 predict the decision to choose to think about the self with 70% accuracy, $p = .01$. In other words, the

brain state a person enters in MPFC/BA10 as soon as their mind is free from external demands predicts whether they next want to think about themselves. But is it the only part of the brain that can do this? We next assessed if MPFC/BA10 predicted self-choices better than patterns from the entire brain. The whole brain pattern during pre-trial rest was able to predict the subsequent decision to choose the self with 77% accuracy, $p = .001$. The higher accuracy for whole brain classification suggests additional brain areas contribute to the bias towards self-focus.

To determine which additional brain regions contribute to the bias towards self-focus, it is important to consider that MPFC/BA10 is one of multiple brain regions comprising the brain's default network and that mental phenomena can arise from distributed patterns of neural activity across interacting brain regions[46–48]. Graph analytic methods indicate that the default network is comprised of three subsystems: the core subsystem (associated with self-reflection, prospection and autobiographical memory), dorsomedial subsystem (associated with social semantic knowledge and inferring mental states), and the medial temporal lobe subsystem (associated with episodic memory, simulation, and relational processing[2,43,49]. As alluded to in its name, the core subsystem is the primary default network subsystem, and MPFC/BA10 is a key node. Thus, while MPFC/BA10 is the region most reliably associated with self-reflection[50], it is possible that distributed patterns of neural activity in the core default network subsystem, including MPFC/BA10, may contribute to the bias towards self-focus.

We therefore repeated the steps for (SVM) classification on a whole-brain activation map masked by the three subsystems of the default network (core subsystem, dMPFC subsystem, and MTL subsystem) as mapped by Yeo et al.[43]. We found multivariate patterns in the core subsystem during pre-trial rest periods predicted the decision to choose to think about the self with 83% accuracy, $p < .001$. The dMPFC subsystem (48% accuracy, $p = .61$) and MTL subsystem (59% accuracy, $p = .17$) both were not able to classify results above chance. These results demonstrate that distributed patterns in the DMN core subsystem during pre-trial rest best predict the subsequent choice to think about the self. The DMN core subsystem classification accuracy is higher than the whole-brain accuracy, suggesting that the whole-brain results are largely driven by the DMN core subsystem. We would expect that even though the whole-brain includes the DMN core,

its accuracy would be lower because of the noise created by other areas of the brain outside of the core system.

We next dug another layer deeper, to understand precisely which DMN core regions besides MPFC/BA10 contribute to its classification power. We performed follow-up analyses with all of the ROI's included in Yeo's DMN core subsystem: right middle temporal gyrus, right superior frontal gyrus, left middle frontal gyrus with left superior frontal gyrus, left angular gyrus, right angular gyrus with right middle temporal gyrus, posterior cingulate/precuneus (PCC), and MPFC/BA10. The only classifiers that were significant with an uncorrected threshold were the PCC (69% accuracy, $p = .02$) in addition to the MPFC/BA10 (70% accuracy, $p = .01$). To determine if one region was contributing more than the other to the core subystem's predictive power, we ran an MVPA analysis with the DMN core subsystem minus the MPFC/BA10 (73% accuracy, $p = .004$), the DMN core subsystem minus the PCC (73% accuracy, $p = .005$), the DMN core subsystem minus the MPFC/BA10 and PCC (59% accuracy, $p = .17$), and the MPFC/BA10 and PCC combined (72% accuracy, $p = .006$). We ran statistical tests on a total of 15 classifier models over the course of this analysis. When we correct for multiple comparisons only the DMN core subsystem (83% accuracy, $p = .0001$) and the whole brain (77% accuracy, $p = .001$) survive the corrected threshold ($p < .05/15$, or .003). These results in combination with the above suggest the MPFC/BA10 and PCC equally contribute to the success of the DMN core classifier accuracy, but the activation pattern of the whole DMN core subsystem is needed for the best classification power.

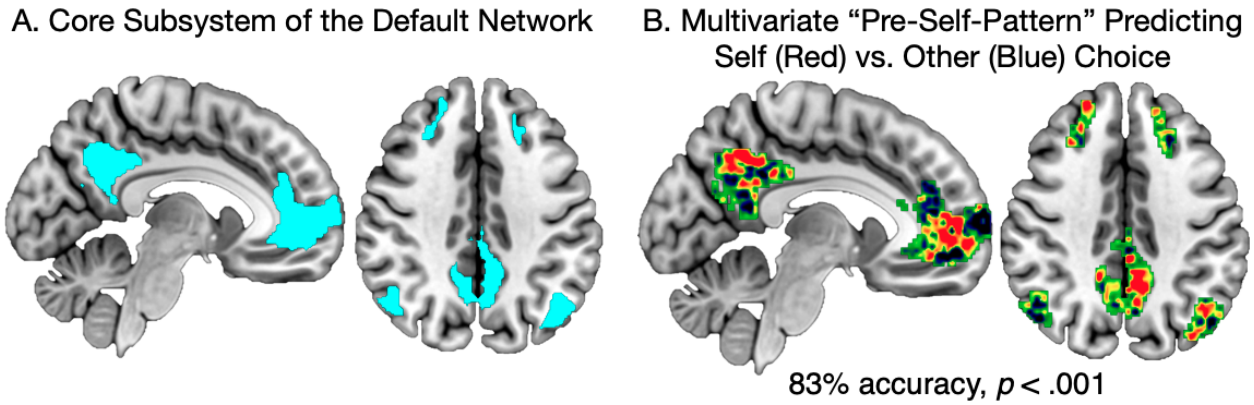

Fig. 3 The DMN core subsystem during pre-trial rest periods predicted the decision to choose to think about the self with 83% accuracy, $p < .001$. (A) depicts the DMN core subsystem. (B) depicts the multivoxel pattern generated by the SVM to differentiate rest

activity that proceeds self-choice rather than other-choice. Red indicates regions where activity predicts self-choice and blue indicates regions where activity predicts other-choice.

A neural signature that predicts self-focus: Evidence from the resting state scan

The Pre-self Pattern in the Default Network Core Subsystem during Long Periods of Rest Predict Self-Reported Self-Focus

So far, we have identified a pre-self pattern which predicts self-focused behavior in a forced choice task. Our next goal was to determine if the pre-self pattern generalizes to predict self-focus in another context—an extended resting state scan. In the beginning of our experiment, participants completed an 8-minute resting state scan and every 2 minutes completed a series of self-reports. They were asked to rate, on a 1-to-5 scale, how much they are thinking about themselves, others, the past, and the future (Figure 1B). Here, we asked: does the presence of the pre-self pattern preferentially predict self-reported self-focus? We performed an instatement analysis–a within-subjects, TR-to-TR multivariate pattern similarity analysis–assessing the similarity between subjects' 1) pre-self pattern and 2) resting state scan pattern in the default network core subsystem. Specifically, each TR of the rest data was masked by the Yeo DMN core subsystem and correlated with the DMN core subsystem pre-self pattern. Correlation values were then averaged for the two-minute rest sections and fisher z-transformed. A linear mixed model assessed how self-reported thought content (self, other, future, and past) as well as the section of rest affected the mean correlation. To ensure that the regressors were not introducing collinearity to the model we ran tolerance and variance inflation factors (VIF). Results indicated that collinearity is not an issue in our model: all tolerance values were over 75% (Self = 82, Other = 90, Future = 77, Past = 88, and Section = 93) and all VIF results were well below the threshold of 4 (Self = 1.2, Other = 1.1, Future = 1.3, Past = 1.1, and Section = 1.1).

A stronger presence of the pre-self pattern corresponded with greater self-reported self-focused thought ($\beta = .19$, standardized $\beta = 0.09$, $t(105.1) = 2.03$, and $p = 0.045$) but not other-focused thought ($\beta = 0.07$, standardized $\beta = 0.08$, $t(104.1) = .85$, and $p = 0.40$), future-focused thought ($\beta = -0.01$, standardized $\beta = 0.08$, $t(92.6) = -0.10$, and $p = 0.92$), past-focused thought ($\beta = 0.05$, standardized $\beta = 0.07$, $t(86.8) = 0.70$, and $p = 0.49$) or the

section of the long rest (β = .06, standardized β = 0.06, $t(80.1) = 0.91$, and $p = 0.37$). Follow-up analyses with the dmPFC and MTL subsystems, as well as the MPFC/BA10 and PCC examined individually, did not produce significant results (β's < 0.14, p's > 0.10). Thus, multivariate patterns in the default network core subsystem—derived from short, jittered rest to predict subsequent choice to think about the self—is also able to predict self-reported, self-focused thought during long rest.

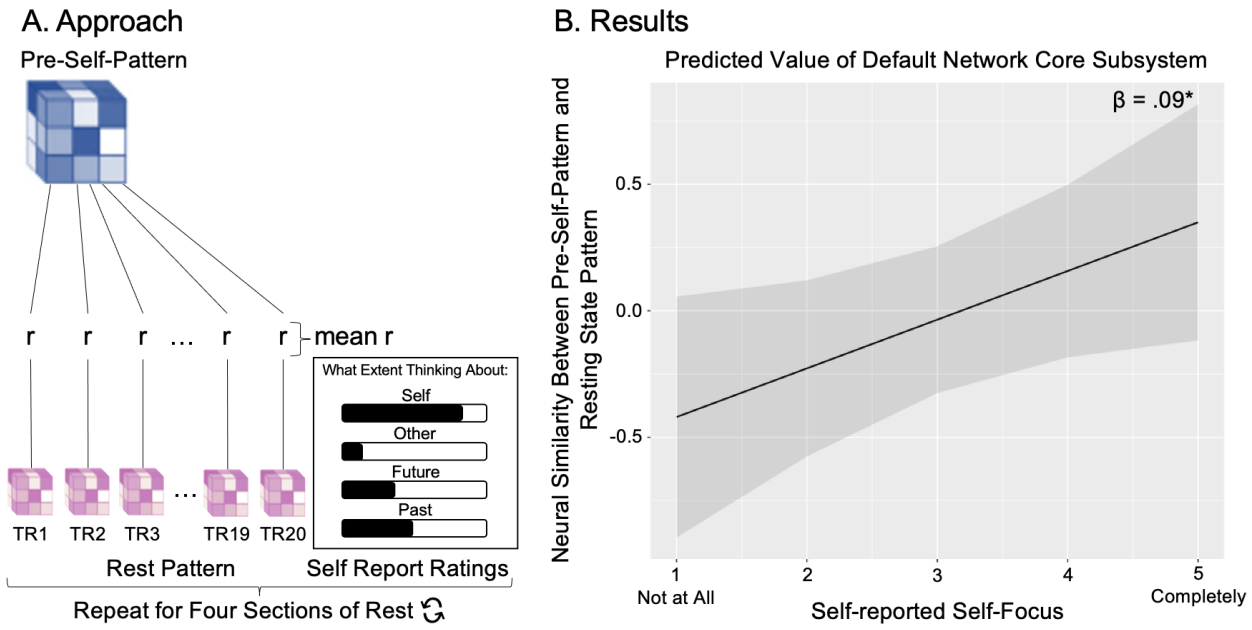

Fig. 4 Applying the pre-self pattern to extended rest with experience sampling. (A) Our analytic approach started with a within-subjects, TR-to-TR multivariate pattern similarity analysis, assessing the similarity between subjects' 1) pre-self pattern and 2) resting state scan pattern in the default network core subsystem. Those correlation values were then averaged for the two-minute rest sections and fisher z-transformed. A linear mixed model assessed how self-reported thought content (self, other, future, and past) as well as the section of rest affected the mean correlation. B) In the results graph, the y-axis displays the normed mean correlation of the long rest neural patterns (in the DMN core) with the multivoxel classifying pre-self pattern generated (in the DMN core). The x-axis is the self-reported self-focus rating that follows each of the two-minute sections of rest from 1 = "not at all" to 5 = "completely". We found that the strength of the DMN core subsystem multivoxel pattern during rest was significantly related to self-reported self-focus (β = .19, standardized β = 0.09, $t(105.1) = 2.03$, $p = 0.045$). In other words, the DMN core

multivariate pattern derived from short, jittered rest to predict subsequent choice to think about the self, is also able to decode self-focused thought during long rest.

The Pre-self Pattern Temporally Predicts the Neural Active Self-Reflection Pattern during Long Periods of Rest

So far, we have assessed whether the pre-self pattern predicts behavioral markers of self-focus: self-focused decisions and self-reported self-focus. Could the pre-self pattern even predict a marker of self-focus that participants do not necessarily report on? To answer this question, we computed, for each subject, a multivariate pattern that reflected their active self-reflection in our final fMRI task, in which they answered questions about themselves such as “am I an ambitious student?”. We then used instatement analysis to test whether the presence of the pre-self pattern precedes the presence of this self-reflection-pattern in the core default network subsystem. We restricted our analyses to the core subsystem because of the accumulating evidence so far that distributed patterns in this network meaningfully predict self-focus. We took each participant’s multivariate self-reflection pattern from the final fMRI task and assessed its neural pattern similarity with each second (i.e., TR) of the resting state scan. We then assessed if resting state neural pattern similarity to the pre-self pattern temporally predicted neural pattern similarity to the self-reflection-pattern. Specifically, linear mixed models assessed if pre-self pattern correlation strength predicted active self-reflection pattern correlation strength 5 TRs later[51]. We selected the length of 5 TRs because the pre-self pattern was generated using a jittered-rest period that was on average 4.5 seconds. Analyses are performed on fisher z-transformed correlation values. It is noteworthy that the pre-self pattern and self-reflection pattern for each subject demonstrated small correlations ($-.03 < r\text{'s} < .08$). We therefore ran the linear mixed models two ways: 1) with the pre-self pattern correlation predicting the self-reflection-pattern 5 TRs later and 2) with the self-reflection pattern predicting the pre-self pattern 5 TRs later. Running both of these analyses helped ensure the small correlation between the two patterns was not driving observed temporal results. In the default network core subsystem, the strength of the pre-self pattern significantly predicted the strength of the active self-reflection pattern 5 TRs later ($\beta = .06$, standardized $\beta = 0.06$, $t(14716) = 2.93$, $p = 0.003$). The opposite was not true; active self-reflection patterns did not significantly predict pre-self patterns 5 TRs later ($\beta = .05$, standardized $\beta = 0.04$, $t(14716) = 1.53$, $p = 0.125$). Thus, multivariate patterns in the default network core

subsystem—derived from short, jittered rest to predict subsequent choices to think about the self—are also able to temporally predict neural indices of active self-reflection during long rest.

We also capitalized on the active self-reflection pattern to test whether the relationship between the pre-self pattern and self-reported self-focus (reported in the section above) is independent of a relationship between the active self-reflection pattern and self-reported self-focus. A linear mixed model, that included both the pre-self pattern and the active self-reflection pattern as independent predictors of self-reported self-focus showed that there was still a significant relationship between pre-self pattern instatement and self-reported self-focus (β = .20, standardized β = 0.10, $t(102.4) = 2.08$, and $p = 0.041$). Indeed, when we looked at variance inflation factors, VIF results were well below the threshold of 4 ('pre-self' pattern = 1.01 and self-reflection pattern = 1.02). This further suggests that the pre-self pattern is not redundant with active self-reflection.

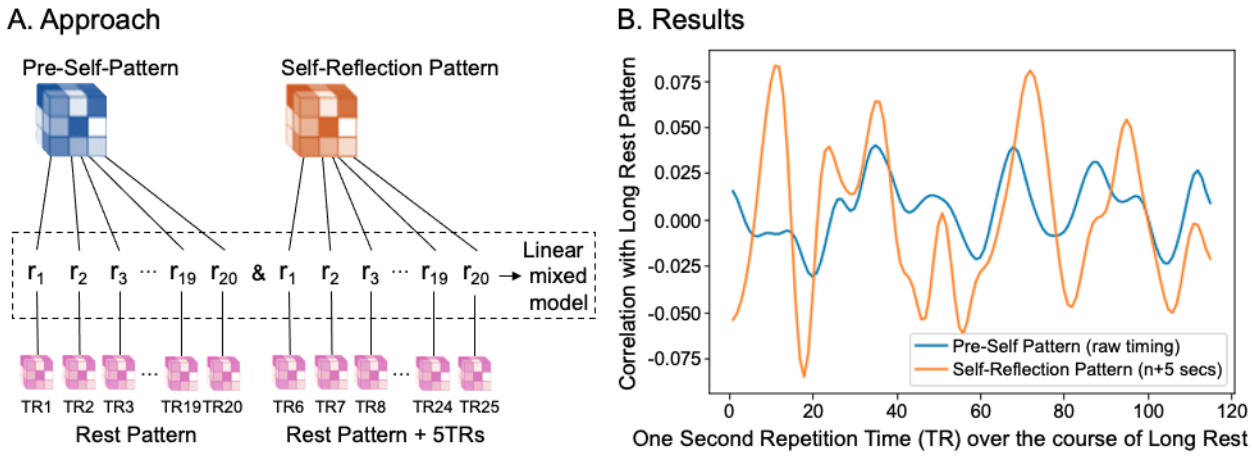

Fig. 5 Relationship between Pre-self Pattern and Active Self-Reflection Pattern Over Long Rest (A) depicts the approach: the pre-self pattern (blue) was correlated with each TR of the resting state scan and the active self-reflection pattern (orange) was also correlated with each TR of the resting state scan. This allowed us to assess whether the presence of the pre-self pattern temporally predicted the presence of the active self-reflection pattern. (B) visualizes the results for a single subject, demonstrating that the presence of the pre-self pattern (blue) predicts the presence of the active self-reflection pattern (orange) 5 seconds later. Note that the self-reflection pattern strength visualized

in orange has been shifted in time (n + 5 seconds) to help visualize its relationship with the pre-self pattern.

A neural signature that predicts self-focus: Evidence from the Human Connectome Project

In the Human Connectome Project Dataset, High Internalizers Rhymically Default to the Pre-self Pattern

The results from our dataset suggest that we can decode the bias towards self-focus–the pre-self pattern predicts self-focused decisions, subjective self-focus during rest, and the presence of active self-reflection neural patterns a few seconds later. Self-focused thought is implicated in mental health conditions, particularly internalizing disorders such as anxiety and depression[52–55]. Could the pre-self pattern be used to predict internalizing scores in an entirely separate sample of participants? If so, this would have great utility: it would suggest we derived a neural marker that, down the line, could be used to identify a person's vulnerability to internalizing conditions. In a first step towards this goal, we took the pre-self pattern, created in our 32 subject dataset, and applied it to a larger dataset to show that the pattern translates not just across tasks but also across datasets and subjects.

We conducted an instatement analysis–a TR-to-TR pattern matching analysis–in the Human Connectome project dataset (N=1086) with our pre-self pattern in the core default network subsystem. Specifically, we assessed the degree of instatement of the pre-self pattern during each TR of the baseline resting state scan. Participants in the Human Connectome dataset each have an internalizing score, which reflects symptoms like anxiety, depression, and withdrawal implicated in maladaptive self-focus. We hypothesized that the extent to which we see the pre-self pattern during rest will be related to participants' internalizing scores. Because maladaptive self-focus is characterized by an incessant focus on the self, we specifically wondered if internalizing predicts the temporal structure of participants' movement into and out of the pre-self pattern.

For every TR, we correlated each subject's internalizing score with their pre-self pattern instatement value. Visual inspection revealed a rhythmic nature to the results of this

analysis, such that greater internalizing appeared to correspond with greater pre-self pattern instatement around the first 150 TRs and rhythmically returned multiple times throughout the resting state scan (Figure 6C). That said, it is difficult to determine from visual inspection alone if the high internalizers drive this temporal structure (or, alternatively, if low internalizers drive the result such that low internalizing is associated with less frequent pre-self pattern instatement). Given that internalizing is associated with maladaptive self-focus, we predicted the result is driven by the high internalizers. To test this prediction explicitly, we completed an inter-subject representational similarity analysis (IS-RSA[41]). The goal of IS-RSA is to determine whether individuals similarly high on a given dimension (here, internalizing) show similar brain responses. Specifically we analyzed the relationship between a subject's internalizing score and the frequencies at which they move in and out of the pre-self pattern.

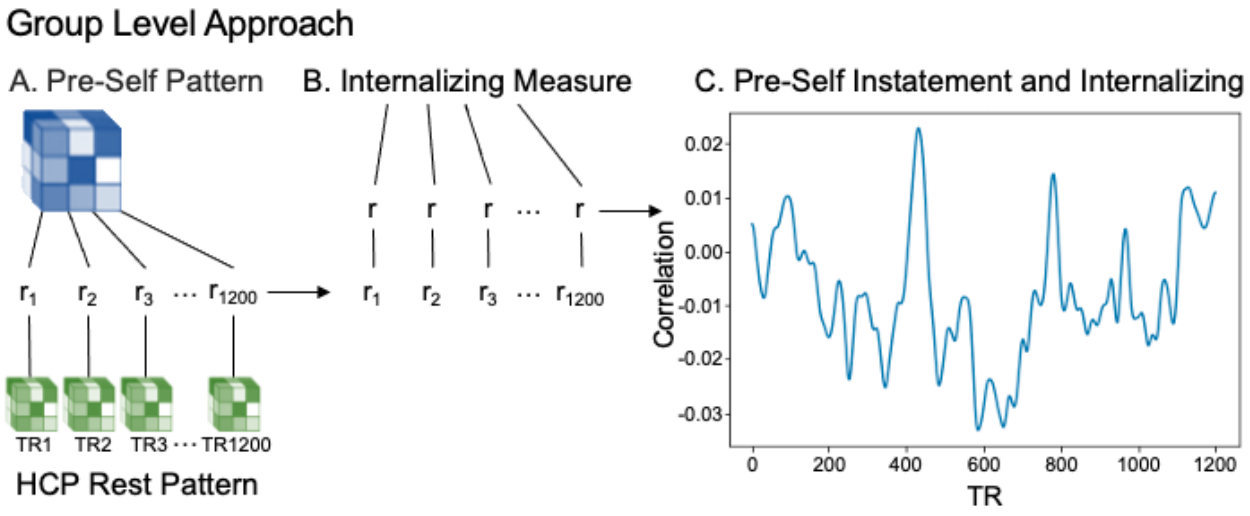

Fig. 6 Human Connectome Project Group Level Analysis Each TR of the 14.5 minutes of rest was masked by the Yeo DMN core subsystem ROI and then correlated with corresponding pre-self pattern. Next we correlated each subjects internalizing score with their pre-self pattern instatement value for each TR. Visual inspection revealed a rhythmic nature to the results of this analysis, such that greater internalizing appeared to correspond with greater pre-self pattern instatement around the first 150 TRs and rhythmically returned multiple times throughout the resting state scan.

We took an "Anna Karenina" model approach to our IS-RSA, named after the opening line of Tolstoy's famous novel, which goes, "All happy families are alike; each unhappy family

is unhappy in its own way"[56]. In our case, we hypothesized that all high internalizers move rhythmically in and out of the pre-self pattern alike, while low internalizers move in and out of the preself pattern in their own unique ways. For each Human Connectome participant, we performed a fourier transformation on their instatement time course–the time course reflecting the presence of the pre-self pattern during each TR of their resting state scan (Figure 7). We wanted to include as many frequencies as made sense given the time dynamics of fMRI and the hemodynamic response. We selected all of the frequencies from the slowest frequency– with a period of the length of the whole 14.5 min rest period– to the frequency with a period of 4 seconds. This generated, for each participant, a vector of sixty-nine frequencies' magnitudes and angles (a total of 138 data points) (Fig 7). We then Pearson correlated each subject's frequency vector with each other to populate an inter-subject frequency similarity matrix.

To connect individual differences in internalizing to neural activity, we first converted subjects' internalizing scores into ranks, making low internalizing subjects ranked low and high internalizing subjects ranked high (range of ranks = 0-1085 for N=1086). Our Anna Karenina model represented the subject pair's similarity of frequency vectors as the mean of the pair's internalizing rank. The higher the pair's internalizing rank, the higher their similarity in pre-self pattern instatement frequency (indicating more normative patterns of movement into and out of the pre-self pattern), and vice-versa. This generated our 1086*1086 inter-subject mean internalizing matrix. See Fig. 7 for a depiction of the Anna Karenina model.

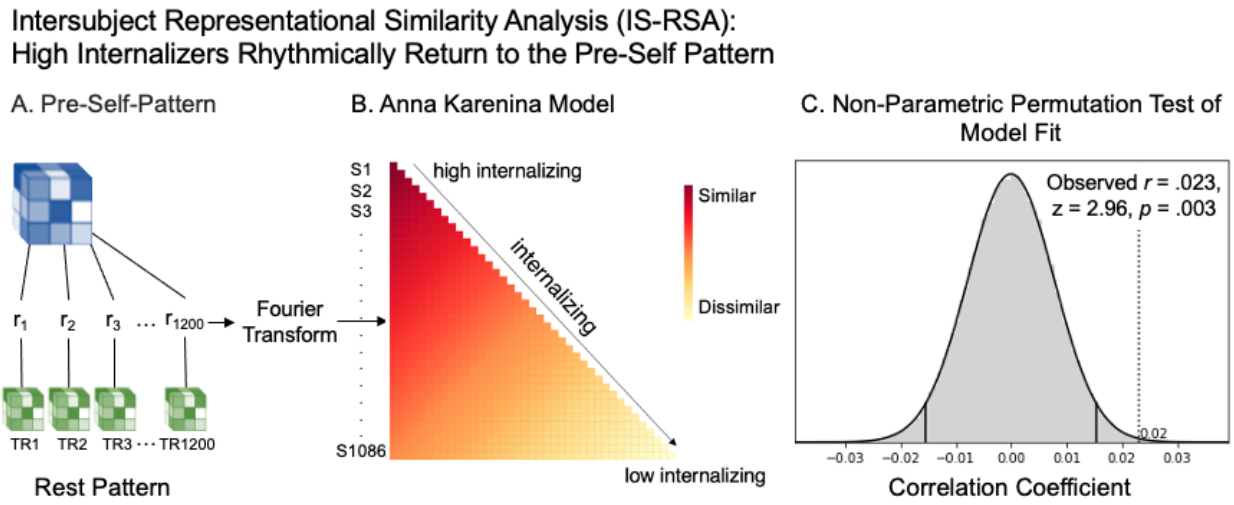

Fig. 7 Intersubject Representational Similarity Analysis (IS-RSA): High Internalizers Rhythmically Return to the Pre-Self Pattern For our IS-RSA we compared the rhythms observed in individuals instatement of the pre-self pattern with a Anna Kerenina model of internalizing such that the higher the pair's internalizing rank, the higher their similarity in pre-self pattern instatement frequency (indicating more normative patterns of movement into and out of the pre-self pattern), and vice-versa. A non-parametric, Mantel permutation test (see Methods) showed the internalizing Anna Karinina model was a significantly good match for the pre-self pattern instatement frequency matrix (r = .023, p = .003).

Finally, these two matrices–the inter-subject mean internalizing and frequency similarity matrices–were correlated and this correlation was statistically tested using a non-parametric, Mantel permutation test (see Methods). The Anna Karenina model was significant (r = .023, p = .003, Mantel permutation test). In other words, people high on internalizing, a clinical variable highly related to maladaptive self-focus, show a similar temporal structure to the presence of the pre-self pattern, repeatedly returning to it throughout an extended resting state scan.

Discussion

We set out to determine if we could find a neural pattern that predicts the bias towards self-focus and this work has done just that. The pre-self pattern in the core default network subsystem decodes the subsequent decision to focus on the self with high accuracy. When we apply it to other outcomes in our dataset, the pre-self pattern in the core default network subsystem also predicts self-reported self-focus during extended rest periods. It is even capable of temporally predicting the presence of a neural pattern capturing active self-reflection during long rest. Finally, when applied in the Human Connectome Project dataset, our pre-self pattern significantly related to internalizing scores in the Human Connectome Project data. Individuals with high internalizing scores moved into and out of the pre-self pattern in similar rhythmic ways during a baseline resting state scan. This result offers a unique way to think about how internalizing relates to brain function–it may systematically warp the temporal trajectories of mind wandering, keeping one's attention incessantly focused on the self.

What is going on–in terms of a psychological process–in participants' brains when we observe the pre-self pattern? Importantly, while the pre-self pattern is modestly correlated with a pattern reflecting active self-reflection (from a task in which participants were instructed to reflect on themselves), it is not redundant with the active self-reflection pattern. Rather, the pre-self pattern temporally predicts the later presence of the active self-reflection pattern and explains unique variance in participants' self-reported self-focus. Whatever psychological process may be captured by our pre-self pattern, it seems to be distinct from active self-reflection.

There may be at least three possible psychological phenomena captured by our pre-self pattern. First, the pre-self pattern may signify a "loading" of the self; a coming online of a self-schema through which to view and interact with the world. Consistent with this view, some posit that we constantly exist in a framework of the self, as we understand our world through our own experiences and embodied sense of self, which we constantly exist in[57]. The second possibility is related to the first: the pre-self pattern could reflect the persistent cognitive accessibility of the self[58–60]. Cognitive category accessibility was first proposed by Bruner[60] and suggests that the attentional framework we are in directly influences what we subsequently perceive or think about. Thus, the pre-self pattern may reflect an attentional state that facilitates quick access to self-knowledge. Third, it is possible the pre-self pattern represents a motivational state. Specifically, participants may be experiencing a temptation or impulse to think about the self. This is an urge that they could potentially resist or succumb to, perhaps with significant implications for mental health, specifically rumination. Future research will help arbitrate between these competing possibilities. Regardless of the potential psychological process reflected in the pre-self pattern, its observation is consistent with predictive coding accounts of brain function, which broadly suggest endogenous, default brain states predict subsequent perception and cognition[21–23].

Part of what makes the pre-self pattern useful is that it predicts self-focus in multiple ways and even across different datasets. This means that it could be utilized as a decoding tool in future research on the role of self-focus in other relevant behaviors. For example, self-focus during conversation corresponds with worse relationship quality[61] and reducing self-concept accessibility has been suggested to help us engage with other people's points-of-view[62]. Yet, objective markers of these phenomena, as well as a clear explanation as

to how they occur, remains to be determined. The pre-self pattern provides a neural signature to better understand the bias towards self-focus in multiple aspects of everyday life.

The pre-self pattern may also be a useful decoding or investigative tool in mental health research. Among Human Connectome Project participants, repeatedly and rhythmically moving into the pre-self pattern during rest corresponded with internalizing symptoms, including depression and anxiety. Rumination–repetitive and recurrent negative thinking about oneself[52]–is thought to play a key role in internalizing disorders[53–55]. Yet, the neural mechanism that explains why and how rumination spontaneously occurs is not fully understood, although past work has associated the default network with depression and anxiety[63,64]. The pre-self pattern may offer key insight into the basic mechanisms underlying rumination. Moreover, the pre-self pattern may be applied to patient populations and/or individuals at risk for developing internalizing disorders to help predict, and perhaps eventually offset, maladaptive self-focus.

Self-focus is a pervasive human phenomenon that plays a vital role in our lives—from detecting our hunger cues to understanding our social standing. However, in its most pernicious forms, self-focus is a risk and maintenance factor for internalizing disorders such as depression and anxiety[14–19]. This paper documents a neural signature–the pre-self pattern in the core subsystem of the default network–that biases people towards self-focus. The pre-self pattern predicts self-focused decisions, subjective self-focus during rest, and the presence of active self-reflection neural patterns a few seconds later. Moreover, rhythmically returning to the pre-self pattern in the core subsystem during rest predicted internalizing scores in a large sample of participants from the Human Connectome Project. With this pre-self pattern in hand, we are one step closer towards understanding why humans can not help but focus on themselves, as well as how this process goes awry in mental health conditions.

Methods

Our Dataset:

Participants

Thirty-two individuals (19 identifying as female, mean age = 22.4 years, SD=4.5 years; 21=Caucasian, 9=Asian, 5=Multiracial, 3=Latina/Hispanic, 1=Pacific Islander, 1=Prefer Not To Say) completed our fMRI study. Participants were eligible to participate if they were safe for MRI scanning (e.g., no metal in their body, not pregnant, not claustrophobic). All participants completed informed consent in accordance with the university institutional review board (IRB).

Procedures

*Resting State Scan***.** The first fMRI run that the participants completed was an 8 minute long resting state scan. The 8 minutes were broken up into four, 2 minute long sections. After each 2 minute section, participants had 32 seconds to rate the extent to which they were thinking about themselves, others, the future, and the past. For the first six participants this rating time was 25 seconds, but it was expanded to make sure participants were able to complete all four of the ratings. This was accounted for in all subsequent analyses by ensuring that all the rating TRs were regressed out in the residual image calculation stage of analysis detailed below. These ratings were made on a continuous scale with 'not at all' on one end, 'completely' on the other end, and 'somewhat' in the middle of the scale. Participants utilized a trackball mouse to make their selection and a value between 1 and 5 and to the 14^{th} decimal point was recorded for each of the categories (self, other, future, and past).

Self-Focus Choice Task. For the self-focus choice task, participants were led to believe that they were selecting the trial types they would receive in a separate task that would follow. In actuality, after completing this choice task, all participants proceeded to get the same task, described below. The choice task started with a pre-trial jittered rest period (2.5–6 sec, mean = 4.5 sec), followed by the choice activity where participants choose who (themselves, a close other (i.e., a self-nominated friend), or a well-known other (i.e., President Biden) they wanted to think about in a later task (5 sec). This design is an adaptation of previous fMRI paradigms used to assess self-reflection[32,66–68]. To make the task more engaging, and to confirm default MPFC/BA10 brain states predict self-focus broadly, as opposed to considering the self along a certain dimension(s), specifically, we included six different categorical dimensions such that participants made choices about who to think about along the following dimensions: social roles (e.g., being a daughter), preferences, physical traits, personality traits, future, and past. Participants made a total

of 108 choices (split into two runs each lasting 11 minutes and 40 seconds), evenly distributed between 18 choices per dimension. After each decision regarding what trial type they would get in a later scanner task, participants completed an attention reorienting trial in which they picked which of two bottom shapes matched the top shape (2–4.5 sec, mean = 3 sec). This attention reorienting task served as a "mental palate cleanse," to get participants' minds off of their last choice before the next jittered rest. Participants were instructed not to actively reflect on the topic and subject that they selected, but instead to make the decision that first came to mind, and next focus on the shape matching task. See Figure 1A for a schematic of the task.

Self-Reflection Task. The last fMRI task that participants completed for the study was an active self-reflection task. In this task participants rated how well an adjective described their personality in different social roles (Friend, Student, Significant Other, Son/Daughter, and Worker) on a scale from 1 to 4 using button boxes. We used the Big 5 list of 100 adjectives[69]. These are adjectives evenly split among the Big 5 (agreeableness, conscientiousness, intellect, emotional stability, and surgency) as well as positive and negative valence. The social role participants reported on shifted every 10 adjectives. We used a change in text color to ensure participants noted this change. The number of adjectives from each of the Big 5, as well as the positive and negative valence, were balanced across the five roles. The rating trials were 4 seconds long and the jittered rest was 1–3 sec, mean = 2 sec. There were two runs of 101 trials each, with a total of 202 trials. We used this task here to generate, for each subject, a neural pattern of their active self-reflection. Other analyses can be run with this task to answer separate theoretical questions, but are outside of the scope of this report and will not be examined here.

Behavioral Analysis

For the Self-Focus Choice Task, we first conducted an analysis of variance (ANOVA) testing whether the number of choices varied across the target factor (self vs. friend vs. Biden). Then, because the interaction was significant, we followed up with paired-sample t-tests to confirm the interaction is driven by more decisions for the self. These paired sample t-tests were performed using the R package *jmv* to compare choice of subject (self, friend, or Biden). Another ANOVA tested whether RT varied across the target factor (self vs. friend vs. Biden). Next, because the results were significant, linear mixed models using the R package *lme4* were constructed to assess how choice of subject (self, friend,

or Biden) affected the response time[51]. We then tested whether choice and, separately, RT, interacts with the dimension (roles, preferences, etc). First, we conducted an ANOVA testing the interaction of the terms, and found that there was a significant interaction with choice but not response time. So because the choice interaction was significant, we followed up with paired sample t-tests on choice of subject (self, friend, or Biden), with data limited to each category (Roles, Preferences, Physical Traits, Personality Traits, Future, and Past). Binomial linear mixed models using the R package *lme4* were constructed to assess if there were changes in choice preferences over the course of the task (i.e., with time). Three models were constructed, one for each of the choices (self, friend, and Biden) with trial number, run number, and the interaction of trial number and run number included as regressors.

fMRI Collection

Brain imaging took place on a Siemens Prisma 3T scanner. Functional runs were acquired using a T2*-weighted echo-planar imaging sequence (2.5-mm voxels, repetition time [TR]= 1,000 ms, time to echo [TE] = 30 ms, 2.5-mm slice thickness, field-of-view [FOV] = 24 cm, matrix = 96, flip angle = 59, and simultaneous multislice = 4). A T2-weighted structural image was acquired coplanar with the functional images (0.9-mm voxels, TR = 2,300 ms, TE = 2.32 ms, 0.9-mm slice thickness, FOV = 24 cm, matrix = 256 256, and flip angle = 8) for the purpose of aligning functional data to brain structure during preprocessing. For the fMRI tasks, sequence optimization was obtained using optimizeXGUI in MatLab[70].

fMRI Preprocessing

For the fMRI dataset we collected, results included in this manuscript come from preprocessing performed using fMRIPrep 20.2.2[71,72] (RRID:SCR_016216), which is based on Nipype 1.6.1[73,74] (RRID:SCR_002502). Per recommendations from the software developers, we report the exact text generated from the boilerplate below.

For each of the 2 BOLD runs found per subject (across all tasks and sessions), 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) are estimated before any spatiotemporal filtering using mcflirt[75] (FSL 5.0.9). A B0-nonuniformity map (or fieldmap)

was estimated based on a phase-difference map calculated with a dual-echo GRE (gradient-recall echo) sequence, processed with a custom workflow of SDCFlows inspired by the epidewarp.fsl script and further improvements in Human Connectome Project Pipelines[76]. The fieldmap was then co-registered to the target EPI (echo-planar imaging) reference run and converted to a displacements field map (amenable to registration tools such as ANTs) with FSL's fugue and other SDCflows tools. Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD reference was then co-registered to the T1w reference using flirt[77] (FSL 5.0.9) with the boundary-based registration[78] cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions[79]) and Jenkinson (relative root mean square displacement between affines[75]). FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al.[79]). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction[80] (CompCor). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al.[80] in that instead of eroding the masks by 2 pixels on BOLD space, the aCompCor masks are subtracted a mask of pixels that likely contain a volume fraction of GM. This mask is obtained by thresholding the corresponding partial volume map at 0.05, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately

within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each[81]. Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym 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 the 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[82]. Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Residual Image Calculation

For the Self-Focus Choice Task, in order to examine neural activity during pre-trial rest that is not biased by neural activity during the choice activity or shape matching activity, we first regressed out the effects of the choice activity and shape matching activity as well as the effects of motion. This step is in line with prior research examining pre-trial neural responses[32,83,84]. All analyses of the pre-trial jitter period were run on the residual images saved from those models. Specifically, residual images were calculated by modeling the choice task and shape-matching task convolved with the canonical hemodynamic response function in a general linear model. This model included nuisance regressors for the six motion parameters (x, y, and z directions and roll, pitch, and yaw rotations), each motion parameter's derivative and square of the derivative, linear drift, and run constants. We additionally regressed out TRs in nonsteady state and TRs that exhibited spikes of motion found from global signal outliers and outliers derived from frame differencing (each 3 SDs). We used a smoothing kernel of 6 on the residual images that were used for the

parametric modulation analysis but no smoothing kernel on the residual images used for the MVPA analysis.

Consistent with prior resting state research[85], we also calculated residual images for the long rest task. In order to examine neural activity during rest that is not biased by neural activity during the self-report rating activity, we first regressed out the effects of the self-report rating activity as well as the effects of motion. All analysis of the long rest periods were run on the residual images saved from those models. Specifically, residual images were calculated by modeling the rating task convolved with the canonical hemodynamic response function in a general linear model. This model included nuisance regressors for the six motion parameters (x, y, and z directions and roll, pitch, and yaw rotations), each motion parameter's derivative and square of the derivative, linear drift, and run constants. We additionally regressed out TRs in nonsteady state and TRs that exhibited spikes of motion found from global signal outliers and outliers derived from frame differencing (each 3 SDs). Finally, we used a smoothing kernel of 6.

Parametric Modulation

Prior work examining pre-trial rest has used parametric modulation analyses to show that greater MPFC/BA10 activity during pre-trial rest accelerates self-reflection[32]. To see if we conceptually replicate this result, tor the self-focus choice task, we performed a parametric modulation analysis, examining Yeo's MPFC ROI[43] specifically, to determine if faster decisions to *choose* to think about the self are preferentially preceded by pre-trial MPFC/BA10 activity. We used nltools[86] in python to create a first-level model (performed on participants' residual images) that comprised a regressor for the rest period before each trial, a regressor of which choice the participant made in the trial that followed the rest period (self, friend, or Biden), and finally, a regressor of the mean centered RT for that trial. We created a contrast of self versus other, where friend and Biden were combined into the other category. We focused on this self versus other contrast so that there were a near equal number of trials in each group (self = 1696, friend = 1127, Biden = 624, other = 1751) and to limit the number of comparisons we made. We then found the mean activation of this contrast in Yeo's MPFC ROI for each subject. Paired sample t-tests were performed on these results using the R package *stats*[87]. A follow-up whole-brain parametric modulation analysis was completed to investigate if any other brain regions demonstrated a parametric effect, and/or if other contrasts showed a parametric

effect. We looked at four contrasts, 1) self vs other (friend and Biden), 2) self vs. friend, 3) self vs. Biden and 4) friend vs. Biden). We then performed a second-level model for each of these contrasts, utilizing nltools[86] in python, with a voxelwise false discovery rate corrected $p < .05$ threshold, contrast 1 k=99 cluster extent threshold, contrast 2 k=109 cluster extent threshold, contrast 3 k=49 cluster extent threshold, contrast 4 k=51 cluster extent threshold as determined by AFNI's 3dClustSim cluster size correction.

Multi-voxel Pattern Analysis (MVPA)

To determine if we can "decode" the bias towards self-focus from pre-trial rest, we performed multi-voxel pattern analysis on a whole brain activation map masked by the MPFC/BA10, the three subsystems of the default mode network (default network core, dMPFC subsystem, and MTL subsystem) as identified by Yeo et al.[43], as well as an unmasked whole brain activation map. These masks were chosen because we were interested in assessing the MPFC/BA10 role's in particular, and the default network's role more broadly in prompting us toward self-focus. We again used the contrast of self versus other, where friend and Biden were combined into the other category. We focused on this self versus other contrast so that there were a near equal number of trials in each group. We trained a linear SVM to discriminate a subsequent choice of self (coded as 1 in the classification) versus other (coded as -1 in the classification) with a k-fold cross-validation procedure[88–90]. In the statistical learning literature[88,91], there are many types of classification algorithms, but they generally perform very similarly on problems such as the one we pursued here. SVM algorithms such as the one we used in this study are the most widely used algorithm for two-choice classification and are robust and reasonably stable in the presence of noisy features.

We computed prediction performance using a 6-fold balanced cross-validation procedure[44,45]. We subdivided the data into 6 separate folds (5-6 participants in each group) and used all of the data except for one fold to train the model and then tested the model using the left-out fold. We then iterated over this process for every possible fold. This process was completed for each of our five neural maps (MPFC/BA10, whole brain, DMN core subsystem, dmPFC subsystem, and MTL subsystem) and an average classification accuracy was calculated for each. A follow up analysis was then done following the same steps with all of the ROI's included in Yeo's DMN core subsystem (right middle temporal gyrus, right superior frontal gyrus, left middle frontal gyrus with left

superior frontal gyrus, left angular gyrus, right angular gyrus with right middle temporal gyrus, and PCC) except for the MPFC/BA10 as that ROI was already analyzed. Another follow up analysis was also done utilizing the same methods with four ROIs, 1) DMN core subsystem minus the MPFC, 2) DMN core subsystem minus the PCC 3) DMN core subsystem minus the MPFC/BA10 and PCC, and 4) the MPFC/BA10 and PCC combined. We ran statistical tests on a total of 15 classifier models over the course of this analysis. To account for issues with multiple comparisons statistical results are only reported as significant if they have $p < .05/15$ or .003.

To test the statistical significance of these results, we generated null distributions for each ROI using 10,000 permutations of a 6-fold support vector machine classification analysis. First, for each subject we relabeled their self and other neural images randomly as self or other. Then we subdivided the data into 6 separate folds (5-6 participants in each group) and used all of the data except for one fold to train the model and then tested the model using the left-out fold. We then iterated over this process for every possible fold. This was repeated 10,000 times and the average classification accuracy of each permutation was used to create a null distribution. Our average classification accuracy result was then compared to this null distribution using nltools[86]. This process was completed for all of the neural maps mentioned above (whole brain, DMN core subsystem, dmPFC subsystem, MTL subsystem, right middle temporal gyrus, right superior frontal gyrus, left middle frontal gyrus with left superior frontal gyrus, left angular gyrus, right angular gyrus with right middle temporal gyrus, PCC, and MPFC/ACC, DMN core subsystem minus the MPFC, DMN core subsystem minus the PC, DMN core subsystem minus the MPFC/BA10 and PCC, and the MPFC/BA10 and PCC combined). Because results from this analysis implicated the DMN core subsystem, all subsequent analyses described below move forward with the DMN core subsystem as the primary mask and follow-up analyses testing for specificity to this subsystem are run with the other Yeo default network subsystems/ROIs.

Self-Reflection Pattern Generation

In order to test whether our pre-self pattern temporally predicts the presence of active self-reflection during passive rest, we needed to create a neural pattern reflecting active self-reflection and subsequently examine instatement of self-reflection neural activity during long rest. To this end, we generated neural patterns for each subject using their self-

reflection task. Specifically, images were calculated by modeling the self-reflection task convolved with the canonical hemodynamic response function in a general linear model. This model included nuisance regressors for the six motion parameters (x, y, and z directions and roll, pitch, and yaw rotations), each motion parameter's derivative and square of the derivative, linear drift, and run constants. We additionally regressed out TRs in nonsteady state and TRs that exhibited spikes of motion found from global signal outliers and outliers derived from frame differencing (each 3 SDs). Finally, we used a smoothing kernel of 6. We created a contrast of self-reflection compared to baseline and saved the resulting image for each subject to use in the subsequent instatement analysis.

Pre-self and Self-reflection Pattern Instatement During Long Rest

Our next approach to assessing our participants' resting state data was to see if the presence of their pre-self pattern predicted their subjective ratings of self-focus during the rest scan. We performed an instatement analysis–a TR-to-TR pattern matching analysis with the pattern generated by the MVPA analysisof the DMN core subsystem. Each TR of the eight minutes of rest was masked by the Yeo DMN core subsystem ROI and then correlated with the DMN core subsystem pre-self pattern. These correlation values were then averaged for the two minutes of rest before each rating and z-scored. Linear mixed models using the R package *lme4* were constructed to assess how self-report of thought content (self, other, future, and past) as well as the section of rest affected the mean correlation[51]. To ensure that the regressors were not introducing collinearity to the model we ran tolerance and variance inflation factors. All VIF results were well below the threshold of 4 (Self = 1.2, Other = 1.1, Future = 1.3, Past = 1.1, and Section = 1.1) and all tolerance values were over 75% (Self = 82, Other = 90, Future = 77, Past = 88, and Section = 93). To follow up on this, analyses were also run with the pre-self patterns generated by the dmPFC subsystem, MTL subsystem, MFPC, and PCC.

Given that we now had, for each subject, their pre-self pattern and active self-reflection pattern (see section above "Self-Reflection Pattern Generation"), we could test whether the presence of the pre-self pattern temporally predicted the presence of the active self-reflection pattern. We performed an instatement analysis–a TR-to-TR pattern matching analysis–with the pattern generated by the MVPA analysis of the DMN core subsystem and the self-reflection pattern generated with the self-reflection task. First, we masked the self-reflection pattern of each subject to the Yeo DMN core subsystem ROI. Then each

TR of the eight minutes of rest was masked by the Yeo DMN core subsystem ROI and correlated with the DMN core subsystem pre-self pattern as well as the masked version of the self-reflection pattern. Linear mixed models using the R package *lme4* were constructed to assess how correlation strength of the pre-self pattern affected the correlation strength of the self-reflection pattern 5 TRs later[51]. We selected the length of 5 TRs because the pre-self pattern was generated using a jittered-rest period that was on average 4.5 seconds. We tested the correlation of the two neural patterns for each subject and did find small (r = .08 to -.03). We therefore ran the above analysis both with pre-self pattern correlation proceeding self-reflection pattern as well as the self-reflection pattern proceeding the pre-self pattern correlation. This enabled us to make sure that the correlation of the two patterns was not the cause of significant temporal results.

Human Connectome Project Dataset:

Participants

To investigate whether the pre-self pattern from our dataset generalizes to predict self-focus in other data, we also used data from the first resting state scan of the Human Connectome Project, hereafter referred to as the Human Connectome Project dataset[65]. The dataset is openly accessible and consists of a large sample of neurotypical individuals. Data from the Human Connectome Project are publicly available in the online Human Connectome Project repository (https://db.humanconnectome.org/; fMRI data are in the subfolders rfMRI_REST1_RL, behavioral data are in the Restricted Data file). We utilized neural and behavioral data from one thousand eighty-six individuals (age 22–37 years, mean age 28.8; 588 female and 498 male; 817=White, 63=Asian/Nat. Hawaiian/Othr Pacific Is., 158=Black or African Am., 2=Am. Indian/Alaskan Nat., 28=More than one, 18=Unknown or Not Reported; 979=Not Hispanic/Latino, 94=Hispanic/Latino, 13=Unknown or Not Reported).

fMRI Collection

We used the first resting state scan from the Human Connectome Project. The resting state scan that we used in our analysis was 14 minutes 33 seconds long and was the first functional scan done on participants' first day in the lab. The fMRI data were acquired using a 3T Skyra scanner with 2 mm isotropic voxels and a TR of 0.72 s (Van Essen et al.[65] for more acquisition details). Each run comprised 1200 scan volumes, and there was

a single run for each participant. We used minimally preprocessed voxelwise fMRI data. See Glasser et al.[76] for further details about data preprocessing.

Residual Image Calculation

Consistent with our own data analysis pipeline and prior resting state research[85], we also calculated residual images for the Human Connectome Project long rest scan. Residual images were calculated by modeling the canonical hemodynamic response function in a general linear model. This model included nuisance regressors for the six motion parameters (x, y, and z directions and roll, pitch, and yaw rotations), each motion parameter's derivative and square of the derivative, linear drift, and run constants. We additionally regressed out TRs in nonsteady state and TRs that exhibited spikes of motion found from global signal outliers and outliers derived from frame differencing (each 3 SDs). Finally, we used a smoothing kernel of 6. All analyses of the Human Connectome Project long rest scan were run on the residual images saved from those models.

Linking Pre-self Pattern Instatement and Internalizing In the Human Connectome Project Dataset

We performed an instatement analysis–a TR-to-TR pattern matching analysis–with the pre-self pattern (e.g. the pattern generated by the MVPA Analysis) in the DMN core subsystem. Each TR of the 14.5 minutes of rest was masked by the Yeo DMN core subsystem ROI and then correlated with corresponding pre-self pattern. Next we correlated each subjects internalizing score with their pre-self pattern instatement value for each TR. Visual inspection revealed a rhythmic nature to the results of this analysis, such that greater internalizing appeared to correspond with greater pre-self pattern instatement around the first 150 TRs and rhythmically returned multiple times throughout the resting state scan (Figure 6). To determine which participants–those with high or low internalizing–were driving this periodic relationship, we decided to complete an inter-subject representational similarity analysis (IS-RSA[41]). Specifically we analyzed the relationship between a subject's internalizing score and the frequencies at which they move in and out of the pre-self pattern.

Inter-subject Representational Similarity Behavioral Model

To connect individual differences in internalizing to neural activity, we first converted subjects' internalizing scores into ranks, making low internalizing subjects ranked low and

high internalizing subjects ranked high (range of ranks = 0-1085 for N=1086). Our Anna Karenina model represented subject pair's similarity of frequency vectors as the mean of the pair's internalizing rank. The higher the pair's internalizing rank, the higher their similarity in pre-self pattern instatement frequency (indicating more normative patterns of movement into and out of the pre-self pattern), and vice-versa. This generated our 1086*1086 inter-subject mean internalizing matrix. See Fig. 7 for a depiction of the Anna Karenina model.

Inter-subject Representational Similarity Neural Model

For each Human Connectome participant, we first carried out an instatement analysis in which we performed a TR-to-TR pattern matching analysis with our pre-self pattern–the pattern generated by the multivoxel pattern analysis of the DMN core subsystem. Each TR of the 14.5 minutes of rest was masked by the Yeo DMN core subsystem ROI and then correlated with the DMN core subsystem pre-self pattern. This created a time course for each subject where each data point indicated the extent to which the participant's neural activity matched our pre-self pattern at that particular time point. Next we performed a fourier transformation on each subject's instatement time course utilizing python's scipy.fft package[92]which generated a vector of 1200 frequencies' magnitudes and angles (a total of 2400 data points). We then performed IS-RSA, testing whether participants high on internalizing showed a similar rhythmic structure to the presence of their pres-self pattern during rest (i.e., an "Anna Karenina" model[41]).

For our IS-RSA we wanted to include as many frequencies as made sense given the time dynamics of fMRI and the hemodynamic response. We selected all of the frequencies from the slowest frequency–with a period of the length of the whole 14.5 min rest period–to the frequency with a period of 4 seconds. This generated, for each participant, a vector of sixty-nine frequencies' magnitudes and angles (a total of 138 data points) (Fig 7). We then Pearson correlated each subject's frequency vector with each other to populate a representational similarity matrix. This frequency similarity matrix was organized as a function of participant's internalizing scores, so that participants higher on internalizing appeared at the top and those lower on internalizing appeared on the bottom.

Inter-Subject Representational Similarity Analysis (IS-RSA) with Mantel tests

Finally, to test our hypothesized link between high internalizing and similar pre-self pattern activity during rest, we first spearman correlated our inter-subject mean internalizing and frequency similarity matrices (specifically, correlating only the lower triangles of these symmetric matrices). We used spearman correlations, as is protocol in the Representational Similarity Analysis literature[93], because the increase in pre-self pattern frequency similarity may not be linear to the increase in the internalizing of a subject-pair. To test our model's statistical significance, we need to account for each subject appearing in the model multiple times—as we compare every subject to every other subject meaning each subject appears N-1 (1086) times in our model. To account for this non-independence in our data, we tested our model with a non-parametric, Mantel permutation test, as was done in previous works[94,95]. Specifically, we randomly shuffled subject's frequency vector identity—each subject's (intact) frequency vector was relabeled with a different subject's identity—100,00 times, each time correlating the resulting simulated frequency similarity matrix with our unshuffled mean internalizing matrix, creating a (null) distribution of IS-RSA (correlation) values. We then quantified the probability that our results were produced by chance by computing the proportion of times our simulated null correlation value exceeded our observed model-data correlation. Finally, to determine statistical significance we compared this probability to a significance threshold of alpha = .05.

Acknowledgements and Funding:
This work was supported by an R01 grant from NIMH awarded to Dr. Meghan L. Meyer. Courtney Jimenez, Sasha Brietzke, Christina Huber, and Megan Hillis assisted with data collection. Luke Chang provided valuable advice on data analysis decisions.

Data Availability:
Data from our experiment can be found here: https://osf.io/8gefx/: and Human Connectome Data that we analyzed can be found here: (https://db.humanconnectome.org/

References

1. Wallace, D. F. *This Is Water: Some Thoughts, Delivered on a Significant Occasion, about Living a Compassionate Life*. (Little, Brown and Company, 2009).
2. Andrews-Hanna, J. R., Reidler, J. S., Huang, C. & Buckner, R. L. Evidence for the Default Network's Role in Spontaneous Cognition. *J. Neurophysiol.* **104**, 322–335 (2010).
3. Song, X. & Wang, X. Mind wandering in Chinese daily lives--an experience sampling study. *PLoS One* **7**, e44423 (2012).
4. Ruby, F. J. M., Smallwood, J., Engen, H. & Singer, T. How self-generated thought shapes mood--the relation between mind-wandering and mood depends on the socio-temporal content of thoughts. *PLoS One* **8**, e77554 (2013).
5. Bischoping, K. Gender differences in conversation topics, 1922–1990. *Sex Roles* **28**, 1–18 (1993).
6. Dunbar, R. I. M., Marriott, A. & Duncan, N. D. C. Human conversational behavior. *Hum. Nat.* **8**, 231–246 (1997).
7. Emler, N. A Social Psychology of Reputation. *European Review of Social Psychology* **1**, 171–193 (1990).
8. Emler, N. Gossip, reputation, and social adaptation. in *Good gossip* 117–138 (University Press of Kansas, 1994).
9. Landis, M. H. & Burtt, H. E. A Study of Conversations. *J. Comp. Psychol.* **4**, 81–89 (1924).
10. Naaman, M., Boase, J. & Lai, C.-H. Is it really about me? message content in social awareness streams. in *Proceedings of the 2010 ACM conference on Computer supported cooperative work* 189–192 (Association for Computing Machinery, 2010).
11. Symons, C. S. & Johnson, B. T. The self-reference effect in memory: a meta-

analysis. *Psychol. Bull.* **121**, 371–394 (1997).

12. Tamir, D. I. & Mitchell, J. P. Anchoring and adjustment during social inferences. *J. Exp. Psychol. Gen.* **142**, 151–162 (2013).
13. Epley, N., Keysar, B., Van Boven, L. & Gilovich, T. Perspective Taking as Egocentric Anchoring and Adjustment. *J. Pers. Soc. Psychol.* **87**, 327–339 (2004).
14. Ingram, R. E. Self-focused attention in clinical disorders: review and a conceptual model. *Psychol. Bull.* **107**, 156–176 (1990).
15. Nolen-Hoeksema, S. Responses to depression and their effects on the duration of depressive episodes. *J. Abnorm. Psychol.* **100**, 569–582 (1991).
16. Nolen-Hoeksema, S. The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *J. Abnorm. Psychol.* **109**, 504–511 (2000).
17. Nolen-Hoeksema, S., Parker, L. E. & Larson, J. Ruminative coping with depressed mood following loss. *J. Pers. Soc. Psychol.* **67**, 92–104 (1994).
18. Just, N. & Alloy, L. B. The response styles theory of depression: tests and an extension of the theory. *J. Abnorm. Psychol.* **106**, 221–229 (1997).
19. Kuehner, C. & Weber, I. Responses to depression in unipolar depressed patients: an investigation of Nolen-Hoeksema's response styles theory. *Psychol. Med.* **29**, 1323–1333 (1999).
20. Denny, B. T., Kober, H., Wager, T. D. & Ochsner, K. N. A Meta-analysis of Functional Neuroimaging Studies of Self- and Other Judgments Reveals a Spatial Gradient for Mentalizing in Medial Prefrontal Cortex. *J. Cogn. Neurosci.* **24**, 1742–1752 (2012).
21. Huang, Y. & Rao, R. P. N. Predictive coding. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 580–593 (2011).
22. Friston, K. Does predictive coding have a future? *Nature neuroscience* vol. 21 1019–1021 (2018).

23. Hutchinson, J. B. & Barrett, L. F. The power of predictions: An emerging paradigm for psychological research. *Curr. Dir. Psychol. Sci.* **28**, 280–291 (2019).
24. Hsieh, P.-J., Colas, J. T. & Kanwisher, N. G. Pre-stimulus pattern of activity in the fusiform face area predicts face percepts during binocular rivalry. *Neuropsychologia* **50**, 522–529 (3/2012).
25. Aly, M. & Turk-Browne, N. B. Attention promotes episodic encoding by stabilizing hippocampal representations. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E420–9 (2016).
26. Park, H. & Rugg, M. D. Prestimulus hippocampal activity predicts later recollection. *Hippocampus* **20**, 24–28 (2010).
27. Addante, R. J., de Chastelaine, M. & Rugg, M. D. Pre-stimulus neural activity predicts successful encoding of inter-item associations. *Neuroimage* **105**, 21–31 (2015).
28. Raichle, M. E. The Brain's Default Mode Network. *Annu. Rev. Neurosci.* **38**, 433–447 (2015).
29. Scalabrini, A. *et al.* The self and its internal thought: In search for a psychological baseline. *Conscious. Cogn.* **97**, 103244 (2022).
30. Feurer, C. *et al.* Resting state functional connectivity correlates of rumination and worry in internalizing psychopathologies. *Depress. Anxiety* **38**, 488–497 (2021).
31. Kucyi, A. *et al.* Enhanced medial prefrontal-default mode network functional connectivity in chronic pain and its association with pain rumination. *J. Neurosci.* **34**, 3969–3975 (2014).
32. Meyer, M. L. & Lieberman, M. D. Why people are always thinking about themselves: Medial prefrontal cortex activity during rest primes self-referential processing. *J. Cogn. Neurosci.* **30**, 714–721 (2018).
33. Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C. & Norman, K. A. Human hippocampal replay during rest prioritizes weakly learned information and

predicts memory performance. *Nat. Commun.* **9**, 3920 (2018).

34. Tambini, A. & Davachi, L. Awake Reactivation of Prior Experiences Consolidates Memories and Biases Cognition. *Trends Cogn. Sci.* **23**, 876–890 (2019).
35. Rogers, T. B., Kuiper, N. & Kirker, W. S. Self-reference and the encoding of personal information. *J. Pers. Soc. Psychol.* (1977) doi:10.1037//0022-3514.35.9.677.
36. Fossati, P. *et al.* In search of the emotional self: an fMRI study using positive and negative emotional words. *Am. J. Psychiatry* **160**, 1938–1945 (2003).
37. Craik, F. I. M. *et al.* In Search of the Self: A Positron Emission Tomography Study. *Psychol. Sci.* **10**, 26–34 (1999).
38. Lieberman, M. D., Jarcho, J. M. & Satpute, A. B. Evidence-based and intuition-based self-knowledge: an FMRI study. *J. Pers. Soc. Psychol.* **87**, 421–435 (2004).
39. Wagner, D. D., Chavez, R. S. & Broom, T. W. Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. *Wiley Interdiscip. Rev. Cogn. Sci.* **10**, e1482 (2019).
40. Wagner, D. D., Haxby, J. V. & Heatherton, T. F. The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdiscip. Rev. Cogn. Sci.* **3**, 451–470 (2012).
41. Finn, E. S. *et al.* Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging. *Neuroimage* **215**, 116828 (2020).
42. Goldberg, L. R. The structure of personality traits: Vertical and horizontal aspects. in *Studying lives through time: Personality and development* 169–188 (American Psychological Association, 1993).
43. Yeo, B. T. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
44. Cohen, J. R. *et al.* Decoding developmental differences and individual variability in

response inhibition through predictive analyses across individuals. *Front. Hum. Neurosci.* **4**, 47 (2010).

45. Izuma, K., Kennedy, K., Fitzjohn, A., Sedikides, C. & Shibata, K. Neural Activity in the Reward-Related Brain Regions Predicts Implicit Self-Esteem: A Novel Validity Test of Psychological Measures Using Neuroimaging. *J. Pers. Soc. Psychol.* (2018) doi:10.1037/pspa0000114.
46. Bressler, S. L. & Menon, V. Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn. Sci.* **14**, 277–290 (2010).
47. Kragel, P. A., Koban, L., Barrett, L. F. & Wager, T. D. Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging. *Neuron* **99**, 257–273 (2018).
48. Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A. & Wager, T. D. A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol.* **13**, e1002180 (2015).
49. Buckner, R. L. & DiNicola, L. M. The brain's default network: updated anatomy, physiology and evolving insights. *Nat. Rev. Neurosci.* **20**, 593–608 (10/2019).
50. Lieberman, M. D., Straccia, M. A., Meyer, M. L., Du, M. & Tan, K. M. Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence. *Neurosci. Biobehav. Rev.* **99**, 311–328 (04/2019).
51. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
52. Watkins, E. R. Constructive and unconstructive repetitive thought. *Psychol. Bull.* **134**, 163–206 (2008).
53. Watkins, E. R. & Roberts, H. Reflecting on rumination: Consequences, causes, mechanisms and treatment of rumination. *Behav. Res. Ther.* **127**, 103573 (2020).

54. du Pont, A., Rhee, S. H., Corley, R. P., Hewitt, J. K. & Friedman, N. P. Rumination and Psychopathology: Are Anger and Depressive Rumination Differentially Associated with Internalizing and Externalizing Psychopathology? *Clin. Psychol. Sci.* **6**, 18–31 (2018).
55. Baez, L. M. *et al.* Identifying real-world affective correlates of cognitive risk factors for internalizing disorders. *Emotion* **23**, 678–687 (2023).
56. Tolstoy, L. *Anna Karenina*. (Penguin, 2004).
57. Christoff, K., Cosmelli, D., Legrand, D. & Thompson, E. Specifying the self for cognitive neuroscience. *Trends Cogn. Sci.* **15**, 104–112 (2011).
58. Libet, B., Gleason, C. A., Wright, E. W. & Pearl, D. K. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain* **106 (Pt 3)**, 623–642 (1983).
59. den Daas, C., Häfner, M. & de Wit, J. Sizing opportunity: Biases in estimates of goal-relevant objects depend on goal congruence. *Soc. Psychol. Personal. Sci.* **4**, 362–368 (2013).
60. Bruner, J. S. On perceptual readiness. *Psychol. Rev.* **64**, 123–152 (1957).
61. Schwartz-Mette, R. A. & Rose, A. J. Conversational Self-Focus in Adolescent Friendships: Observational Assessment of an Interpersonal Process and Relations with Internalizing Symptoms and Friendship Quality. *Journal of Social and Clinical Psychology* vol. 28 1263–1297 Preprint at https://doi.org/10.1521/jscp.2009.28.10.1263 (2009).
62. Kaufman, G. F. & Libby, L. K. Changing beliefs and behavior through experience-taking. *J. Pers. Soc. Psychol.* **103**, 1–19 (2012).
63. Coutinho, J. F. *et al.* Default mode network dissociation in depressive and anxiety states. *Brain Imaging Behav.* **10**, 147–157 (2016).
64. Sheline, Y. I. *et al.* The default mode network and self-referential processes in

depression. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 1942–1947 (2009).

65. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62–79 (2013).
66. Jenkins, A. & Mitchell, J. Medial prefrontal cortex subserves diverse forms of self-reflection. *Soc. Neurosci.* **6**, 211–218 (2011).
67. Kelley, W. *et al.* Finding the self? An event-related fMRI Study. *J. Cogn. Neurosci.* **14**, 785–794 (2002).
68. Zhu, Y., Li, Z., Fan, J. & Han, S. Neural Basis of Cultural Influence on Self-Representation. *Neuroimage* **34**, 1310–1316 (2007).
69. Goldberg, L. R. The Structure of Phenotypic Personality Traits. *Am. Psychol.* 9 (1993).
70. Spunt, B. *easy-optimize-x: Formal Release for Archiving on Zenodo*. (Zenodo, 2016).
71. Esteban, O. *et al.* fmriprep. *Softw. Pract. Exp.* (2018).
72. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
73. Gorgolewski, K. *et al.* Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* **5**, 13 (2011).
74. Gorgolewski, K. J. *et al.* Nipype. *Softw. Pract. Exp.* (2018).
75. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
76. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).
77. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5**, 143–156 (2001).

78. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* **48**, 63–72 (2009).
79. Power, J. D. *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**, 320–341 (2014).
80. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**, 90–101 (2007).
81. Satterthwaite, T. D. *et al.* 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 (2013).
82. Lanczos, C. Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis* **1**, 76–85 (1964).
83. Inagaki, T. K. & Meyer, M. L. Individual differences in resting-state connectivity and giving social support: implications for health. *Soc. Cogn. Affect. Neurosci.* **15**, 1076–1085 (2020).
84. Spunt, B., Meyer, M. & Lieberman, M. The Default Mode of Human Brain Function Primes the Intentional Stance. *J. Cogn. Neurosci.* **27**, 1–9 (2015).
85. Meyer, M. L., Davachi, L., Ochsner, K. N. & Lieberman, M. D. Evidence That Default Network Connectivity During Rest Consolidates Social Information. *Cereb. Cortex* **29**, 1910–1920 (2019).
86. Luke Chang, Eshin Jolly, Jin Hyun Cheong, Anton Burnashev, Andy Chen, Marissa Clark, Seth Frey, Paxton Fitzpatrick. *cosanlab/nltools*. doi:10.5281/zenodo.10059041.
87. R Core Team. R: A Language and Environment for Statistical Computing. Preprint at https://www.R-project.org/ (2022).
88. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data*

Mining, Inference, and Prediction. (Springer, 2008).

89. Wager, T. D. *et al.* An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).
90. Woo, C.-W. *et al.* Separate neural representations for physical pain and social rejection. *Nat. Commun.* **5**, 5380 (2014).
91. Yu, H. *et al.* A Generalizable Multivariate Brain Pattern for Interpersonal Guilt. *Cereb. Cortex* **30**, 3558–3572 (2020).
92. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
93. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, (2008).
94. Finn, E. S., Corlett, P. R., Chen, G., Bandettini, P. A. & Constable, R. T. Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nat. Commun.* **9**, 2043 (2018).
95. Iyer, S., Collier, E., Broom, T. W., Finn, E. S. & Meyer, M. L. Individuals who see the good in the bad engage distinctive default network coordination during post-encoding rest. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2306295121 (2024).