



HAL
open science

Brain neuromarkers predict self- and other-related mentalizing across adult, clinical, and developmental samples

Dorukhan Açıl, Jessica Andrews-Hanna, Marina Lopez-Sola, Mariët van Buuren, Lydia Krabbendam, Liwen Zhang, Lisette van der Meer, Paola Fuentes-Claramonte, Edith Pomarol-Clotet, Raymond Salvador, et al.

► **To cite this version:**

Dorukhan Açıl, Jessica Andrews-Hanna, Marina Lopez-Sola, Mariët van Buuren, Lydia Krabbendam, et al.. Brain neuromarkers predict self- and other-related mentalizing across adult, clinical, and developmental samples. 2025. <hal-05136713>

HAL Id: hal-05136713

<https://hal.science/hal-05136713v1>

Preprint submitted on 30 Jun 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC 4.0 - Attribution - Non-commercial use - International License

1 **Brain neuromarkers predict self- and other-related mentalizing across**
2 **adult, clinical, and developmental samples**

3
4 Dorukhan Açıl^{1,2,3}, Jessica R. Andrews-Hanna^{4,5}, Marina Lopez-Sola^{6,7}, Mariët van
5 Buuren⁸, Lydia Krabbendam⁸, Liwen Zhang⁹, Lisette van der Meer^{10,11}, Paola Fuentes-
6 Claramonte^{12,13}, Edith Pomarol-Clotet^{12,13}, Raymond Salvador^{12,13}, Martin Debbané^{14,15},
7 Pascal Vrticka¹⁶, Patrik Vuilleumier¹⁷, David A. Sbarra⁴, Andrea M. Coppola⁴, Lars O.
8 White³, Tor D. Wager¹⁸, & Leonie Koban^{19*}

9
10
11 ¹ Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

12 ² Department of Child and Adolescent Psychiatry, Psychotherapy, and Psychosomatics,
13 Leipzig University, Leipzig, Germany

14 ³ Department of Clinical Child and Adolescent Psychology and Psychotherapy, University
15 of Bremen, Bremen, Germany

16 ⁴ Department of Psychology, University of Arizona, Tucson, Arizona, USA

17 ⁵ Cognitive Science, University of Arizona, Tucson, Arizona, USA

18 ⁶ Department of Medicine, School of Medicine and Health Sciences, Institute of
19 Neurosciences, University of Barcelona, Spain

20 ⁷ Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

21 ⁸ Department of Clinical, Neuro and Developmental Psychology, Faculty of Behavioral
22 and Movement Sciences, Institute for Brain and Behavior Amsterdam, Vrije Universiteit
23 Amsterdam, The Netherlands

24 ⁹ Institute for Medical Imaging Technology, Ruijin Hospital, Shanghai Jiao Tong
25 University School of Medicine, Shanghai, China

26 ¹⁰ Department of Clinical and Developmental Neuropsychology, University of Groningen,
27 Groningen, The Netherlands

28 ¹¹ Department of Psychiatric Rehabilitation, Lentis Zuidlaren, The Netherlands

29 ¹² FIDMAG Germanes Hospitalàries Research Foundation, Barcelona, Spain

30 ¹³ Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM) ISCIII,
31 Barcelona, Spain

32 ¹⁴ Developmental Clinical Psychology Research Unit, Faculty of Psychology and
33 Educational Sciences, University of Geneva, Switzerland

34 ¹⁵ Research Department of Clinical, Educational and Health Psychology, University
35 College London, United Kingdom

36 ¹⁶ Department of Psychology, University of Essex, Colchester, United Kingdom

37 ¹⁷ Laboratory of Behavioural Neurology and Imaging of Cognition, Department of
38 Neuroscience, University Medical Center, University of Geneva, Geneva, Switzerland

39 ¹⁸ Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH,
40 USA

41 ¹⁹ Lyon Neuroscience Research Center (CRNL), CNRS, Inserm, Université Claude
42 Bernard Lyon 1, Bron France

43
44
45 Running Head: BRAIN SIGNATURES OF MENTALIZING

46
47
48 *Please address correspondence to:

49 Dr. Leonie Koban, CRNL, Institut des Épilepsies IDEE, 59 Boulevard Pinel, 69500 Bron,
50 France; email: leonie.koban@cnrns.fr

51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74

Abstract

Human social interactions rely on the ability to reflect on one's own and others' internal states and traits—a psychological process known as mentalizing. Impaired or altered self- and other-related mentalizing is a hallmark of multiple psychiatric and neurodevelopmental conditions. Yet, replicable and easily testable brain markers of mentalizing have so far been lacking. Here, we apply an interpretable machine learning approach to multiple datasets (total $N=281$) to train and validate fMRI brain signatures that predict 1) mentalizing about the self, 2) mentalizing about another person, and 3) both types of mentalizing. We test their generalizability across healthy adults, adolescents, and adults diagnosed with schizophrenia and bipolar disorder. The classifier trained across both types of mentalizing showed 98% predictive accuracy in independent validation datasets. Self-mentalizing and other-mentalizing classifiers had positive weights in anterior/medial and posterior/lateral brain areas respectively, with accuracy rates of 82% and 77% for out-of-sample prediction. Classifier patterns across cohorts revealed better self/other separation in 1) healthy adults compared to individuals with schizophrenia and 2) with increasing age in adolescence. Together, our findings reveal consistent and separable neural patterns subserving mentalizing about self and others—present at least from the age of adolescence and functionally altered in severe neuropsychiatric disorders. These mentalizing signatures hold promise as mechanistic neuromarkers to measure social-cognitive processes in different contexts and clinical conditions.

75

Author Note

76

77 This work was funded by a Starting Grant from the European Research Council (ERC,
78 101041087) to LKo, a German Academic Exchange Service (DAAD) doctoral grant and
79 a Network of European Neuroscience Schools (NENS) exchange fellowship to DA, an
80 R01 grant from the U.S. National Institutes of Mental Health (R01MH125414-01) to JAH
81 and DAS, a Junior Leader Fellowship from "la Caixa" Foundation
82 (LCF/BQ/PR22/11920017) to PFC, a Consolidator Grant from the European Research
83 Council (ERC, 648082) to LKr, R37 and R01 support from the U.S. National Institutes of
84 Mental Health (R37MH076136 to TDW, MH116026 to TDW and L. Chang [PI],
85 R01EB026549 to TDW and M. Lindquist [MPis]). LvdM acknowledges a European
86 Science Foundation EURYI grant (044035001) that funded her doctoral studies (PI: A.
87 Aleman). Views and opinions expressed are however those of the authors only and do not
88 necessarily reflect those of the European Union or the European Research Council.
89 Neither the European Union nor the granting authority can be held responsible for them.
90 The funders had no role in study design, data analysis, manuscript preparation, or
91 publication decisions. Matlab code for all analyses is available at: canlab.github.org.

92

Introduction

93

Mentalizing—representing and inferring the psychological states of oneself and others—is a fundamental process for adaptive navigation through the social world (Moore & Frye, 1991; Wellman, 2014). Delineating the brain systems involved in mentalizing about self and others is important for understanding brain health and dysfunction, as atypical mentalizing patterns underlie many neurodevelopmental and psychiatric conditions (Brüne & Brüne-Cohrs, 2006; Debbané et al., 2016; Gray et al., 2011; Luyten et al., 2020; Sharp, 2006; Sloover et al., 2022; Johnson et al., 2022).

100

Many studies have examined the neural correlates of mentalizing, suggesting an interplay of different brain regions, including medial prefrontal cortex (mPFC), temporoparietal junction (TPJ), and precuneus (Frith & Frith, 2006; Saxe, 2006; Schurz et al., 2021; Van Overwalle & Baetens, 2009; Yang et al., 2015). However, predictive brain measures of mentalizing that can be applied to individuals to decode the degree of self-related and other-related processing are still lacking. Most cognitive and affective processes cannot be captured by activity in individual brain regions, as they are reflected in patterns of brain activity distributed across multiple regions and systems, which can be harnessed in decoding (Kragel et al., 2018; Rosenberg et al., 2018). For instance, recent work demonstrates that distributed brain activity and connectivity patterns as indexed by fMRI enables us to decode the intensity of pain (Wager et al., 2013), drug and food craving (Koban et al., 2023), sustained attention (Rosenberg et al., 2016), depressive rumination (Kim et al., 2023), and clinically relevant behaviors and outcomes (Gabrieli et al., 2015). These predictive brain activity patterns or ‘*brain signatures*’ are multivariate models that utilize data across the whole brain to make formally testable population-level predictions across subjects and datasets (Kragel et al., 2018; Woo et al., 2017). The predictions address the involvement and/or the intensity of a mental process. Here, we apply this ‘signature’ approach to predict mentalizing about oneself or other people.

118

Mentalizing, as a hidden state, would arguably benefit from such an approach. Thinking about self and others is inherently multilayered and multidimensional (Schurz et al., 2021; Tamir & Thornton, 2018; Qin et al., 2020). Recent conceptualizations of self- and social-cognition point to multiple dimensions, such as about self versus others, affective versus cognitive (Corradi-Dell’Acqua et al., 2014; Luyten et al., 2020; Schurz et al., 2021) and multiple layers, such as observing behaviours, and processing at the state and trait levels (Tamir & Thornton, 2018). Moreover, social cognition terms, including but not limited to mentalizing (e.g., empathy, perspective taking) suffer from inconsistent

125

126 usage in the literature and lack of consensual definitions (Quesque et al., 2024).
127 Developing brain signatures of mentalizing could potentially help account for this
128 heterogeneity, offering the possibility of testing whether the proposed subdimensions of
129 mentalizing are subserved by dissociable neurobiological patterns. In turn, these
130 signatures carry the potential for validating the multiple facets of mentalizing (Kragel et al.,
131 2018). However, it remains unclear—especially considering the complexity of the
132 mentalizing construct—whether distributed neural patterns can reliably predict mentalizing
133 using brain images.

134 A central question is whether mentalizing about oneself and mentalizing about
135 others (self- and other-mentalizing hereafter) are reliably distinguishable based on brain
136 activity. Recent electrophysiological evidence suggests that self- and other-mentalizing
137 activate overlapping cortical areas following a similar temporal sequence:
138 Temporoparietal junction (TPJ), medial temporal gyrus (MTG)/temporal poles (TP),
139 precuneus/posterior cingulate cortex (PCC), medial prefrontal cortex (mPFC) in a roughly
140 posterior to anterior temporal order (Tan et al., 2022; Wang et al., 2021). Conversely,
141 differences in brain activity between self- and other-mentalizing have also been observed.
142 In mPFC, more dorsal areas coincide with other-related processing and ventral areas with
143 self-related processing, with research initially supporting a linear (Denny et al., 2012) but
144 more recently a curvilinear (Parelman et al., 2021) ventral-to-dorsal gradient for self versus
145 other within mPFC. Further, self-referential thought typically elicits activation in cortical
146 midline structures, such as anterior cingulate cortex (ACC), subcortical areas, including
147 thalamus, striatum and caudate nucleus, and, to a lesser extent, in insula, temporal poles
148 and ventrolateral prefrontal cortex (vlPFC; Fossati et al., 2003; Denny et al. 2012; Murray
149 et al., 2014; Northoff et al., 2006; Parelman et al., 2021; Van der Meer et al., 2010). In
150 contrast, other-referential thought typically elicits activation in TPJ, middle and superior
151 temporal gyri extending to temporal poles (Arioli et al., 2023; Frith & Frith, 2006; Parelman
152 et al., 2021; Saxe, 2006; Tamir et al., 2016; Wagner et al., 2019), and to a lesser degree
153 in supplementary motor area (SMA), left inferior and medial frontal gyri (IFG/MFG) and
154 medial orbitofrontal cortex (Arioli et al., 2021; Murray et al., 2014). Collectively, the
155 literature remains inconsistent regarding the separation between self- and other-
156 mentalizing and it is unclear whether generalizable brain models of self- versus other-
157 mentalizing can be identified.

158 To address these gaps, we leverage fMRI and machine learning to develop three
159 distinct brain signatures, 1) the Mentalizing Signature (MS) for mentalizing overall (i.e.,

160 thinking about either the self or another person versus non-mentalizing control conditions),
161 2) the Self-Referential Signature (Self-RS) to detect specifically self-related thought (here
162 referred to as “self-mentalizing”), and 3) the Other-Referential Signature (Other-RS) to
163 detect other-related thought (here referred to as “other-mentalizing”). This allowed us to
164 test whether dissociable and generalizable neural patterns underlie distinct dimensions of
165 mentalizing.

166 To this end, we pooled data across nine cohorts from six independent fMRI
167 studies. All studies included a mentalizing or a social-cognition task with three conditions:
168 self-processing, other-related processing, and a non-social control condition. The training
169 and validation datasets used variants of a standard trait-evaluation task, which involved
170 reflecting on personality traits/statements describing self or another person. Conceivably,
171 personality traits represent the enduring mental states that can be inferred from social
172 perceptual systems (Molapour et al., 2021) across multiple observations (Schurz et al.,
173 2021; Tamir & Thornton, 2018). Trait representations can be used to predict others’ future
174 behavior and become the basis of the “mental-self” (Qin et al., 2020). Moreover, trait
175 evaluations are one of the few mentalizing tasks used in the literature to include a self-
176 condition, as other commonly used tasks (e.g., mental attribution, false-belief, perspective-
177 taking tasks) only permit mentalizing about others. Thus, trait evaluation is an ideally
178 suited task to study a key mentalizing process, namely representing stable psychological
179 states of self and others.

180 We first used standard machine learning algorithms—support vector machines—
181 to develop and cross-validate multivariate classifiers (signatures) of mentalizing in a
182 sample of healthy adults. In a second step, we then further validated these signatures in
183 seven completely independent test datasets from different laboratories, countries, and
184 scanners, and with different sample characteristics (healthy, adolescent, and clinical
185 populations), allowing us to test their generalizability and predictive validity in adolescent
186 and clinical populations. Third, we tested the signatures in yet another independent
187 dataset that used a different social cognition task to see whether mentalizing signatures
188 would generalize to other contexts in which mentalizing is not directly instructed but likely
189 to implicitly occur. Finally, we assessed whether local patterns of brain activity in several
190 regions of interest (ROIs) contain sufficient information to predict self- and other-referential
191 mentalizing. Together, the results of these analyses inform us about the functional neural
192 organization of self- versus other-related mentalizing and provide us with distributed brain

193 signatures of mentalizing that can be used as brain targets for monitoring and intervening
194 on mentalizing-related brain processes in future studies.

195

196

Results

197 Data overview

198 The study included a total of 904 contrast images from 281 participants and nine
199 independent cohorts, including four samples of healthy adults ($n=118$), two samples of
200 healthy adolescents ($n=105$; $M_{age}=12.9$ and 16), and three samples of adults with clinical
201 diagnosis of either schizophrenia ($n=40$) or bipolar disorder ($n=18$; see Fig. 1a and Table
202 S1). Thus, this study combined six independent studies that constituted the training
203 dataset, validation datasets, and the extension dataset (see Fig 1a). The training and
204 validation datasets included participants completing variants of a self- and other-referential
205 judgement task. The extension dataset included images of participants performing a social
206 feedback task. All tasks included three conditions, namely a Self-condition, an Other-
207 condition, and a non-social Control condition. Contrast images were computed for each
208 condition (versus implicit baseline) and rescaled using L2-norm to standardize the scale
209 of beta weights across participants, studies, and scanners.

210

211 Training and cross-validation results

212 The training dataset (Study 1a) consisted of $n=21$ adult participants who completed
213 a trait-evaluation task using a block design (similar to Kelley et al., 2002; see Fig. 1c). In
214 each block, participants were presented with several trait adjectives: In blocks of the Self-
215 reflection condition, they were asked to rate the degree to which each trait adjective
216 described themselves. In the Other-reflection condition, they had to rate how much each
217 adjective described another person—a confederate with whom the subject had previously
218 interacted during a decision-making task (Koban et al., 2014). In the non-mentalizing
219 Control condition, participants indicated the number of syllables in the trait adjectives. To
220 reduce the influence of non-mentalizing-related brain regions (e.g., visual cortex) and
221 opportunistic classification based on features not related to mentalizing, we used an
222 inclusive mask of brain regions related to social processing (see Fig. 1B).

223 We used support vector machines (SVM) using default parameters (to avoid
224 overfitting) and 10-fold cross-validation (Scheinost et al., 2019) to train three distinct
225 mentalizing signatures using the masked contrasts images from the training dataset. The
226 Self-Referential Signature (Self-RS) was trained to detect Self-reflection (versus the two-

227 remaining conditions), the Other-Referential Signature (Other-RS) to detect Other-
228 reflection (versus the two-remaining conditions), and the Mentalizing Signature (MS) to
229 detect both types of mentalizing versus the Control condition (see Fig. 1c).

230 All three signatures showed excellent cross-validated (out-of-sample) prediction
231 accuracy (100% accuracy in two-alternative forced-choice tests, $p < .001$, averaged
232 *Cohen's d* for Self-RS = 3.18, for Other-RS = 2.45, for MS = 4.92). To identify which voxels
233 contributed most reliably to the mentalizing signatures, we used bootstrapping (5000
234 samples) to obtain one p-value per voxel and displayed the thresholded weight maps
235 using false discovery rate (FDR) correction at $q < .05$ and cluster extent $k > 10$ voxels.
236 Brain regions with significant positive voxel weights for the Self-RS (see Fig. 2) included
237 the vmPFC, dmPFC, frontal eye fields, ventral ACC, frontal operculum, anterior insula,
238 thalamus, caudate nucleus (see Table S2). For the Other-RS, significant positive weights
239 were found in left vIPFC, left STS, bilateral TPJ, and precuneus/PCC (see Fig. 2 and Table
240 S3). The MS had significant positive clusters in mPFC (both dorsal and medial), bilateral
241 vIPFC, dorsal ACC (dACC), frontal operculum, bilateral SMA, bilateral MTG, bilateral TP,
242 left STS, middle cingulate gyrus (MCC), bilateral posterior cingulate cortex (PCC),
243 precuneus, bilateral angular gyrus/temporoparietal junction, and subcortical areas,
244 including right caudate nucleus, left anterior insula (AI), bilateral thalamus (see Fig. 2 and
245 Table S4).

246 For completeness, we trained a Self-versus-Other Classifier following the same
247 training and validation pipelines (see Supplementary Figure 1). For the purposes of
248 simplicity and because the pattern of results is in line with the main three signatures, we
249 report the results pertaining to this signature only in the supplementary figure.

250

251 **Validation in independent samples**

252 Next, we validated the three brain signatures on several completely independent
253 studies: two samples of healthy adults (Study 4a & 5a), two adolescent samples (Study 2
254 & 3), one cohort of participants with bipolar disorder (Study 5c), and two cohorts of
255 participants with schizophrenia (Study 4b & 5b). All datasets included comparable trait
256 evaluation tasks and fMRI block designs with three conditions, with some small variations
257 in task designs between studies (see Fig 1a). To obtain pattern expression values, we
258 computed the matrix dot product between the mentalizing signatures and each subject-
259 level contrast images from these datasets, yielding one scalar value per individual contrast
260 image and per signature (see Fig. 1c). These pattern expression values were then used

261 to test the predictions of the mentalizing signatures. The average prediction accuracy in
 262 two-choice tests, across all independent validation datasets, was 81.52% for the Self-RS
 263 (+/- 6.17% average STE; significant in 10 out of 14 validation tests [7 samples*2
 264 comparisons]), 77.25% for the Other-RS (+/- 6.02% average STE; significant in 12 out of
 265 14 tests), and 97.87% for the MS (+/- 01.79% average STE; significant in all 14 tests),
 266 suggesting overall high prediction accuracy of the three signatures, even in new samples,
 267 although with some variations between datasets (see Fig. 2 & Table S5).

268

269 ***Better self/other separation in healthy adults compared to participants with***
 270 ***schizophrenia***

271 The results above show that the signatures significantly predicted mentalizing in
 272 most samples, including clinical samples. Yet, schizophrenia in particular is often
 273 associated with impaired social cognition and altered self-perception (Bora et al., 2009;
 274 Sprong et al., 2007). Thus, we next tested how well the signatures separated self- from
 275 other-related mentalizing in two of the validation studies (Study 4 & Study 5) that included
 276 both participants with schizophrenia (total n=40) and matched healthy participants (total
 277 n = 48). As expected, both the Self-RS ($M_{HC} = .32$, $M_{SCZ} = .12$; $\beta = .21$, $STE = .07$,
 278 $CI = [.07, .35]$, $p = .004$) and the Other-RS ($M_{HC} = .43$, $M_{SCZ} = .24$; $\beta = .19$, $STE = .07$,
 279 $CI = [.04, .33]$, $p = .01$) showed better discrimination between self- and other-referential
 280 mentalizing (i.e., a greater positive difference between Self-RS responses in the Self
 281 compared to the Other condition, and a greater difference between Other-RS responses
 282 for the Other compared to the Self condition) in healthy adults, compared to adults with
 283 schizophrenia. No group differences were found in the correct discrimination of
 284 mentalizing versus Control by the MS ($p = .28$, see Fig. 3). These results indicate that,
 285 compared to healthy adults, participants with schizophrenia have less differentiated brain
 286 patterns between self- and other-related thought, indicating the potential clinical utility of
 287 the signatures.

288 For completeness, we also compared the discrimination of self- versus other-
 289 related mentalizing activity in participants with bipolar disorder ($n=18$) versus healthy
 290 adults ($n=15$) using data from Study 5. However, participants with bipolar disorder did not
 291 differ significantly from controls for any of the signatures.

292

293

294

295 ***Better self-other separation with increasing age in adolescents***

296 Next, we explored potential developmental differences in the classifier
 297 performances, by testing whether the ability of the classifiers to separate self- from other-
 298 related mentalizing depended on the age of the participants in the two adolescent samples
 299 (Studies 2 and 3, total $N = 105$). We combined data from Study 2 (age 12-18 years; M_{age}
 300 = 16) and Study 3 (age 11-14 years; $M_{age} = 12.9$). We found that (controlling for study)
 301 older adolescents had better self/other separation both for the Self-RS ($\beta = .08$, $STE =$
 302 $.03$, $CI = .012$ to $.018$, $p = .01$) and the Other-RS ($\beta = .06$, $STE = .03$, $CI = .001$ to $.12$, p
 303 = $.047$; see Fig. 4). In contrast, there were no significant associations between age and
 304 performance of the MS. These results suggest that, with increasing age, adolescents'
 305 brain activity becomes more differentiated for self- versus other-related mentalizing.

306

307 **Testing the mentalizing signatures in a social feedback task**

308 So far, we have shown that the signatures performed well in several independent
 309 datasets from different labs, but all using a similar explicit mentalizing task. In order to
 310 examine how the signatures would respond to other types of social tasks, especially those
 311 without any explicit mentalizing demands, we next tested their performance in an fMRI
 312 dataset (Study 6) of $n = 49$ romantic partners performing a social feedback task.
 313 Participants were instructed that their task was to read other participants' likability ratings
 314 about themselves and about their romantic partners. Thus, participants viewed the positive
 315 and negative ratings targeted to themselves (Self-feedback) and to their partners (Partner-
 316 feedback). Participants were informed that their partners would see the same material,
 317 establishing a sense of shared experience. As the control condition, participants viewed
 318 others' feedback without any ratings, which was ostensibly due to technical errors.

319 Here, we tested whether the signatures' responses paralleled the target of the
 320 feedback conditions. For instance, to provide evidence of successful extension, the Self-
 321 RS should show higher pattern expression values in the self-feedback condition as
 322 opposed to other two conditions. As expected, the Self-RS had different response levels
 323 across task conditions in favor of the Self-condition, $F(2,96) = 14.214$, $p < .001$, $\eta_p^2 = .23$
 324 (Fig. 5). Bonferroni-corrected pairwise comparisons showed that the Self-feedback
 325 condition ($M = .16$) had significantly higher Self-RS expression scores than both the
 326 Partner-feedback ($M = -.09$, $p = .01$) and Control ($M = -.36$, $p < .001$) conditions. Additionally,
 327 Self-RS also produced higher pattern expression values for the Partner-feedback ($M = -$
 328 $.09$) compared to Control condition ($M = -.36$, $p = .04$). Similarly, the MS successfully

329 discriminated both feedback conditions against Control condition, $F(2,96)= 14.711$,
330 $p<.001$, $\eta_p^2 = .24$ (Fig. 5). Bonferroni-corrected pairwise comparisons indicated that both
331 Self-feedback ($M= .17$, $p<.001$) and Partner-feedback ($M= .12$, $p=.002$) conditions had
332 significantly higher pattern expression scores than the Control condition ($M= -.17$).
333 However, the Other-RS did not produce any significant differences between task
334 conditions, $F(2,96) = 2.003$, $p = .14$ (Fig. 5). Together this suggests a modulation of the
335 MS and the Self-RS (but not the Other-RS) even in a task where mentalizing was not
336 directly instructed, in line with the idea that feedback to oneself or one's partner should
337 lead to more mentalizing related brain activity.

338

339 **Local patterns of self- and other-related mentalizing**

340 Finally, we trained local classifiers in regions of interest (ROI) associated with
341 mentalizing to gain further insight into how mentalizing-related information is processed
342 locally and which regions can predict self- versus other-related mentalizing. The whole
343 brain patterns indicate which regions have the most reliable positive and negative
344 contributions to different forms of mentalizing, but they do not necessarily show which
345 regions are *not* involved and do not inform us whether local patterns alone can predict the
346 target of mentalizing (self or other). Training and testing classifiers for brain regions
347 implicated in the literature can inform us in this regard.

348 To this end, we included ten ROIs (see Fig. 6) previously associated with
349 mentalizing as shown in an automated term-based meta-analysis (NeuroSynth, Yarkoni,
350 2011): the medial prefrontal cortex (mPFC), two clusters in anterior middle temporal gyrus
351 (aMTG) bilaterally, two clusters in temporoparietal junction (TPJ) bilaterally, a cluster
352 covering precuneus and posterior cingulate cortex (PRE/PCC), left supplementary motor
353 area (SMA), and three clusters in Cerebellum. We trained four types of classifiers in each
354 of these ten ROIs. In other words, we trained each ROI to perform the four following
355 classification tasks separately: 1) Self-mentalizing (versus the two remaining conditions);
356 2) Other-mentalizing (versus the two remaining conditions), 3) both mentalizing conditions
357 (self and other) against Control condition, and 4) Self-mentalizing against Other-
358 mentalizing. The ROI classifiers were trained and cross-validated in our training dataset
359 (Study 1a; $n=21$) and tested in the combined sample ($n=211$) of participants from all
360 validation datasets (including all healthy, developmental, and clinical cohorts; for detailed
361 results see Supplementary Table 6).

362 As expected, all ten regions showed excellent classification accuracy as the
 363 general mentalizing classifiers in the training dataset, and all predicted mentalizing
 364 successfully in the validation dataset (see blue bars in Fig. 6). However, not all regions
 365 could significantly differentiate Self-reflection and Other-reflection conditions. While
 366 mPFC, aMTG, TPJ, and PRE/PCC were capable of differentiating self and other
 367 conditions, the Cerebellum clusters and left SMA failed at this task (see orange bars in
 368 Fig. 6). When trained for self-mentalizing (see red bars in Fig. 6), the right aMTG cluster
 369 could not capture any consistent configurations for Self-mentalizing. While TPJ subregions
 370 were successful here against control conditions, they could not predict Self- against Other-
 371 mentalizing, implying a lack of exclusive neural patterns for self-mentalizing within TPJ
 372 subregions. Finally, when trained for Other-mentalizing, the Cerebellum subregions and
 373 left SMA were incapable of picking up consistent patterns (see violet bars in Fig. 6).

374 Taken together, our analyses suggest that mPFC (including both vmPFC and
 375 dmPFC), TPJ, aMTG, and Precuneus/PCC clusters encode both self- and other-
 376 mentalizing, and that the left SMA and Cerebellum clusters most likely are involved in
 377 domain-general operations during mentalizing. Besides, our findings suggest that TPJ and
 378 aMTG are more strongly associated with other-mentalizing, while mPFC and
 379 Precuneus/PCC clusters are unique in the sense that they are consistently involved in
 380 both self- and other-related mentalizing.

381

382

Discussion

383 Mentalizing—reflecting about others’ and one’s own internal states—is a
 384 fundamental capacity required to function in a social world. Transdiagnostically,
 385 mentalizing deficits typify an array of psychopathological conditions (Brüne & Brüne-
 386 Cohrs, 2006; Debbane et al., 2016; Gray et al., 2011; Luyten et al., 2020; Sharp, 2006;
 387 Sloover et al., 2022; Johnson et al., 2022). Here, we developed a novel brain signature of
 388 mentalizing in general (the ‘*Mentalizing Signature*’ or *MS*), as well as specific self- and
 389 other-related mentalizing signatures (*Self-RS* and *Other-RS*), combining data from nine
 390 independent cohorts ($N=281$). The three signatures showed good to excellent
 391 classification accuracy in independent data and provide several novel insights into the
 392 brain circuits of mentalizing. First, mentalizing appears to coincide with a specific pattern
 393 of brain activity, with reliably dissociable activation patterns for self- and other-related
 394 mentalizing, based on both whole-brain activity as well as within key nodes of the social
 395 brain, especially mPFC, precuneus/PCC, TPJ and aMTG. Second, the signatures

396 significantly predicted mentalizing not only in healthy adults, but also in adolescents and
397 in individuals with schizophrenia and with bipolar disorder, demonstrating its utility across
398 different (including clinical) samples. Third, our results point to the relevance of these
399 mentalizing signatures for a range of different research questions in social, developmental,
400 and clinical neuroscience, by showing (i) an engagement of the MS in an independent
401 social feedback task, (ii) that the differentiation of self- and other-related mentalizing
402 increases as individuals transition from adolescence to adulthood, and (iii) that it is less
403 pronounced in individuals with schizophrenia compared to healthy controls. Thus, our data
404 provide us with clearly defined *brain models* (Kragel, et al., 2018), applicable across
405 contexts to assess the engagement of mentalizing-related brain regions in different
406 experimental conditions and their alteration in clinical and neurodevelopmental conditions.

407 The Mentalizing Signature (MS) successfully predicted mentalizing across
408 different tasks, extending to a social feedback task that did not include any explicit
409 mentalizing instruction. The regions that contributed to the MS most positively included
410 mPFC, precuneus, temporoparietal junction, superior temporal sulcus, and many other
411 regions previously associated with mentalizing (Frith & Frith, 2006; Oosterwijk et al., 2017;
412 Van Overwalle & Baetens, 2009; Tamir et al., 2016; Tan et al., 2022), providing convincing
413 face validity of this signature. Its exceedingly high predictive performance in validation
414 datasets implies that mentalizing recruits consistent and reliable neural configurations in
415 the brain, not only in healthy controls, but also clinical cohorts and developing adolescents.
416 Interestingly, the MS also extended to other tasks predicting social feedback conditions
417 that potentially evoke implicit mentalizing. This indicates that implicit and explicit forms of
418 mentalizing largely share a common neural basis, in keeping with previous research (Van
419 Overwalle & Vandekerckhove, 2013). Thus, subnetworks of mentalizing (implicit-explicit)
420 may be building on a common anatomical architecture that is here covered by the MS.

421 Recent work proposes that self- and other-mentalizing activates common neural
422 constellations or representations and recruits largely overlapping regions (Oosterwijk et
423 al., 2017; Tan et al., 2022; Wang et al., 2021). In contrast, our findings suggest that, while
424 they share a large common core (as evident in the MS), they can also be distinguished
425 reliably based on both whole brain as well as local activation patterns. The Self-RS had
426 positive weights in anterior mentalizing regions around cortical midline structures, such as
427 mPFC, ACC, thalamus, caudate nucleus, frontal eye fields, and insula, extending to the
428 frontal operculum. This pattern is in line with previous meta-analyses (Denny et al., 2012;
429 Murray et al., 2014; van der Meer et al., 2010). The mPFC has been consistently related

430 to mentalizing, and especially its ventral part with self-related processing (Fossati et al.,
431 2003; Kelley et al., 2002; Heatherton et al., 2006; Koban et al., 2021; Wagner et al., 2019).
432 The ventral-to-dorsal linear (Denny et al., 2012) and curvilinear gradient (Pareiman et al.,
433 2021) hypotheses propose that mPFC is involved in both self- and other-person related
434 processing. Accordingly, our results show that mPFC is involved in both self- and other-
435 related mentalizing with different neural configurations, and that the information encoded
436 in mPFC seems to predict self-mentalizing more consistently. Besides, self-processing
437 recruited reward-processing circuits (e.g., striatum), in line with the idea that introspection
438 about oneself is intrinsically rewarding (Chavez et al., 2017; Northoff & Hayes, 2011; Tamir
439 & Mitchell, 2012). The involvement of subcortical areas (e.g., thalamus) and cortical areas
440 associated with interoception (e.g., insula) is in line with the previous work suggesting that
441 one has access to affective/physiological information during self-mentalizing which are
442 less present during mentalizing about others (Maresh & Andrews-Hanna, 2021) and that
443 interoceptive information might be an important contribution to one's sense of self (Babo-
444 Rebelo & Tallon-Baudry, 2018; Garfinkel et al., 2013; Qin et al., 2020).

445 The Other-RS subsumed more posterior and lateral regions, such as TPJ,
446 PCC/Precuneus, STS, and vIPFC, resonating with previous findings (Arioli et al., 2023;
447 Frith & Frith, 2006; Van Overwalle & Baetens, 2009; Pareiman et al., 2021; Saxe, 2006;
448 Tamir et al., 2016; Wagner et al., 2019). Interestingly, our results concerning the
449 precuneus/PCC do not align with the known subdivisions of the default-mode (Andrews-
450 Hanna et al., 2010) or mentalizing networks (Wang et al., 2021). This region is thought to
451 be part of the midline core of the default-mode network (Andrews-Hanna et al., 2010) and
452 of the medial subsystem of the mentalizing network (Wang et al., 2021) with stronger
453 associations with self-processing. However, here in our results, while the precuneus/PCC
454 cluster also contains information about self-mentalizing, it appears more strongly
455 associated with other-related mentalizing, along with the areas that are part of the lateral
456 subdivisions of both networks (e.g., TPJ, temporal lobes). A potential explanation here
457 could be the involvement of precuneus in mental orientation/imagery (Peer et al., 2015;
458 Schurz et al., 2014), which is required in a trait-evaluation task.

459 The mentalizing signatures generalized to data from different types of cohorts,
460 including healthy adults, schizophrenia and bipolar samples, and adolescents. This shows
461 that the overall neurobiological configurations of self- and other-related mentalizing are
462 comparable across these populations and that the signatures have utility in different types
463 of study populations. However, the Self-RS and Other-RS signatures also showed

464 sensitivity to clinical status, since self-versus-other differentiation in the pattern responses
465 was significantly less pronounced for both patterns in participants with schizophrenia
466 compared to matched healthy controls. This finding is noteworthy and in line with clinical
467 observations of alterations in self-perception, mentalizing, and the ability to discriminate
468 between self- and other-generated thought (Bora et al., 2009; Potvin et al., 2019; Sprong
469 et al., 2007; van der Meer et al., 2010). The responses in the bipolar sample did not differ
470 significantly from those of the healthy adults but given the limited sample size of the bipolar
471 group in the present analysis, future work is needed to assess more fine-grained (and
472 potentially context-dependent) differences in mentalizing responses in bipolar disorder, as
473 well as in other psychiatric and neurodevelopmental disorders with mentalizing deficits.

474 In the two adolescent samples (aged 11 to 18 years), greater age was associated
475 with better self/other differentiation of both the Self-RS and the Other-RS. In other words,
476 responses of these two brain signatures to self- and other-related mentalizing were more
477 different from each other in older adolescents. Considering that the signatures were
478 developed using an adult dataset, this reflects an ongoing development of the mentalizing
479 neurobiology (Crone & Fuligni, 2020; Fehlbauer et al., 2022) to become more adult-like
480 and more differentiated for different targets of mentalization, during the age span covered
481 in our study (11-18 years). Future studies could test the role of pubertal development
482 instead of chronological age, and test the signatures in younger children, and those with
483 developmental disorders.

484 We note that the signatures' classification performance in the training and cross-
485 validation sample is consistently higher than in the independent validation datasets. This
486 could potentially owe to leakage of information across the different subjects in the training
487 dataset, who all performed the same task in the same scanner and experimental settings.
488 Alternatively, it might reflect the fact that the Other-condition in the mentalizing task was
489 about a relatively unfamiliar other person (a confederate). This may lead to a greater
490 difference in self- versus other-related processing compared to studies in which
491 mentalizing was about a familiar or close person—which is often more closely related to
492 the self and likely also engages self-referential processing. This could potentially also
493 explain why the Other-classifier was less successful in separating other-related from self-
494 related feedback in the sample of romantic couples (Study 6), where participants likely
495 consider their partners as an extension of themselves. The signatures may also pick up
496 unrelated information in control conditions, which may lead to failed predictions as
497 observed in Study 4 for Self-RS and Study 3 for Other-RS.

498 Future studies may address other important research questions that were beyond
 499 the scope of the present study, such as testing whether other dimensions of mentalizing
 500 (e.g., cognitive versus affective; Luyten et al., 2020; Schurz et al., 2021) or different types
 501 of mental state content (i.e., beliefs, preferences; Defendini & Jenkins, 2023) involve
 502 distinct mentalizing signatures as well. Recent work (Kim et al., 2024; Kim Lux et al., 2022)
 503 investigated valence and self-relevance as two key components of the internal thought. It
 504 is an open question how the brain integrates valence with different targets during
 505 mentalizing. Finally, future research can also build on this work by testing the signatures
 506 in other mentalizing tasks (e.g., false-belief, emotion imagery), on a greater variety of
 507 targets (see Courtney & Meyer, 2020), and in other related mental processes (e.g.,
 508 autobiographical memory retrieval).

509 In conclusion, we trained and validated three whole-brain signatures that predict
 510 self-related, other-related, and both types of mentalizing in multiple independent datasets
 511 that used different variants of a standard mentalizing task in different and diverse
 512 populations. Our findings imply that self- and other-mentalizing use largely dissociable
 513 neural mechanisms that build on a foundational overall mentalizing capacity. Indeed, the
 514 three mentalizing signatures possess potential for use as mechanistic neural markers of
 515 mentalizing about self and others, for example by testing how brain-circuits of mentalizing
 516 are engaged or modulated by different types of experimental conditions, and how they
 517 might be altered in different psychiatric and neurodevelopmental populations.

518

519

520

Methods

521 Participants

522 The present study pooled data from a total of $N=281$ participants (119 females and
 523 162 males, $M_{age}=25.5$, $SD_{age}=12.9$) from nine cohorts participating in six independent
 524 studies (see Fig. 1a and Table S1). These nine cohorts comprised four samples of healthy
 525 (neurotypical) adult samples, two samples of healthy adolescents, and three samples of
 526 adults with a diagnosed psychiatric condition. While most of the datasets have been
 527 previously published separately, the analyses reported here were not published
 528 previously, and the six studies have not been previously combined.

529 The training sample (Study 1; Koban et al., 2014) consisted of $n=21$ healthy adults
 530 (10 women and 11 men, $M_{age} = 23.5$) who were recruited at the University of Geneva,

531 Switzerland. One additional participant with structural abnormalities in the brain was
532 excluded from the original study and the present analysis.

533 Study 2 (Debbane et al., 2017) included $n=44$ healthy adolescents (23 females
534 and 21 males, $M_{age} = 16$, $SD_{age} = 1.86$, Age range = 12.01 to 18.84) who were recruited
535 from secondary schools in Geneva, Switzerland. In the original study, one additional
536 subject was excluded due to structural abnormalities in the brain, three due to
537 incompleteness of the paradigm, one due to signs of substance use, and five due to
538 excessive movement.

539 Study 3 (van Buuren et al., 2020) included $n=61$ healthy adolescents (27 females
540 and 34 males, $M_{age} = 12.9$, $SD_{age} = 0.43$, Age range = 11.61 to 14.22) who were recruited
541 for a longitudinal project from secondary schools in the Netherlands. An additional 18
542 participants were excluded from the original study due to excessive movement, incorrect
543 task completion, or measurement errors.

544 Study 4 (Fuentes-Claramonte et al., 2019; Fuentes-Claramonte et al., 2020)
545 included $n=33$ healthy adults (14 women and 19 men, $M_{age} = 41.7$) and $n=23$ adults with
546 schizophrenia (7 women and 16 men, $M_{age} = 37$). The two cohorts were matched on age,
547 sex, and a measure of general intelligence. The patients were recruited from a local
548 psychiatric hospital in Barcelona, Spain based on diagnostic interviews. The
549 schizophrenia diagnosis was confirmed using the Structured Clinical Interview for DSM
550 Disorders (SCID; First, 2015).

551 Study 5 (Zhang et al., 2015) included three types of cohorts: Healthy adults ($n=15$,
552 6 women and 9 men, $M_{age} = 33.3$), adults with schizophrenia (SZ, $n=17$, 6 women and 11
553 men, $M_{age} = 35.5$), and adults with bipolar disorder (BD, $n=18$, 9 women and 9 men, M_{age}
554 = 40.3). The clinical cohorts were recruited from a local hospital in the north of the
555 Netherlands. The diagnoses of the patients were confirmed using the Mini International
556 Neuropsychiatric Interview-Plus 5.0.0 (MINI-Plus; Sheehan et al., 1998). All BD patients
557 were chosen among those who had a history of at least one psychotic episode. All three
558 cohorts were matched with one another on age, sex, level of education, and a measure of
559 general intelligence. The SZ and BP patients were additionally matched on the level of
560 cognitive and clinical insight as measured by the Schedule of Assessment of Insight-
561 Expanded version (SAI-E, clinical insight; Kemp & David, 1997) and the Beck Cognitive
562 Insight Scale (BCIS, cognitive insight; Beck et al., 2004).

563 Study 6 included 56 healthy adults in romantic relationships recruited from the
564 Tucson, Arizona community and surrounding areas. Seven participants were excluded

565 from the study due to missing or inadequate imaging data, yielding a final sample size of
566 $n=49$ (26 women and 23 men, $M_{age} = 22.6$). Community members were eligible to
567 participate if they had been in a romantic relationship for at least six months, had no
568 contraindications for MRI scanning, and did not meet criteria for active psychosis or mania
569 at the time of screening. Both members of the couple completed all components of the
570 study including the social feedback fMRI task.

571 All participants gave written informed consent and were compensated for their
572 participation via monetary means or gifts. All studies were approved by the respective
573 institutional ethics committees. For additional information, please refer to the original
574 studies and Supplementary Table 1.

575

576 **Tasks**

577 ***Training dataset***

578 We used a trait evaluation task, adapted from Kelley et al. (2002) as the training
579 mentalizing task. Participants were asked to rate the extent to which certain personality
580 adjectives (e.g., 'talkative', 'daring') described themselves (Self-condition) or a same-sex
581 confederate (Other-condition), with whom they thought they were interacting in a
582 preceding decision-making task (Koban, et al., 2014). In the Control condition, they were
583 asked to count the syllables for each trait adjective.

584 Participants met the confederate in person at the beginning of the experimental
585 session where they were briefed that one of them would be tested in the scanner, the
586 other one in a separate room, and that they would interact via computer interface. They
587 first participated in a social decision-making task resembling a serial dictator paradigm
588 whereby the participant was given the choice to share or keep the resources with the
589 confederate in a series of trials (Koban, et al., 2014). After this task, participants were
590 introduced to the trait-evaluation task and were asked to rate their co-player.

591 The task consisted of 150 trials in total presented in 30 blocks (ten per condition).
592 Each block contained five trials and lasted for 20 seconds, with an inter-block-interval
593 (fixation cross) of 8s. Half of the blocks only contained negative and the other half only
594 positive adjectives. Each word appeared once for each condition (50 adjectives were used
595 in total).

596 Each trial started with a 3s cue above the fixation cross reminding the condition
597 (self, other, or syllables) which was followed by the presentation of the adjective and a
598 Likert scale (1-4) on which the participant was expected to select using a button press

599 device. A word was presented every 4 seconds. If the participant made a choice sooner,
600 the word would disappear for the remaining time of the 4 seconds before the next trial.
601 The order of the blocks and words displayed in each block was randomized.

602

603 ***Validation datasets***

604 All validation datasets included similar mentalizing tasks with the same three
605 conditions (Self, Other, Control) as the training task and utilized a block design (for an
606 overview, see Fig. 1a and Table S1). The main difference between validation tasks was
607 the stimuli that were presented: Studies 2 and 3 used trait adjectives, Study 4 (with two
608 cohorts) used trait statements, Study 5 (with three cohorts) used a mix of trait and physical
609 statements, and Study 6 used positive and negative feedback about the likeability of
610 participants' romantic partners, information which was assumed by participants to be
611 accessible to their partner. The studies also used a variety of *others* in the Other-condition:
612 Best/close friend (Study 2), similar and dissimilar classmate (Study 3), relative or close
613 friend (Study 5), romantic partner (Study 6), and acquaintance (Study 4). Studies 2 and 5
614 collected responses using Likert scales and Studies 3 and 4 asked for binary choices
615 (yes/no). Study 6 was primarily a passive viewing task, promoting spontaneous empathy
616 when feedback was directed to participants' romantic partners (or self). Finally, the Control
617 conditions also varied between studies. Studies 4 and 5 presented general knowledge
618 statements; Study 3 asked participants to search for the letter in words; Study 2 asked
619 them to count the syllables in words; and Study 6 presented no feedback during control
620 trials.

621

622 **fMRI data acquisition and pre-processing**

623 ***Training dataset***

624 The training MRI images were acquired on a 3T Magnetom TIM Trio whole-body
625 scanner (Siemens, Germany) with the product 12-channel head coil. A T1-weighted
626 MPRAGE sequence (TR = 1900ms, TI = 900 ms, TE = 2.27 ms, voxel size 1 x 1 x 1 mm)
627 was used to acquire structural anatomical images. Functional images were obtained using
628 a standard T2-weighted echo-planar imaging sequence (2D-EP, TR = 2100 ms, TE = 30
629 ms, flip angle 80°, voxel size 3.2 x 3.2 x 3.2 mm) that scanned the whole brain in 36
630 sequential slices. An automated shimming procedure was included to minimize magnetic
631 field inhomogeneities.

632 SPM8 (Wellcome Department of Imaging Neuroscience, UCL, London, UK) and
633 Matlab® (The MathWorks Inc.) were used for image preprocessing and first-level analysis.
634 A standard preprocessing pipeline was performed, that included spatial realignment and
635 reslicing, coregistration, unified segmentation and normalization to the standard Montreal
636 Neurological Institute (MNI) echo planar imaging template (voxel size: 2 mm³), and finally
637 spatial smoothing using an 8 mm Full Width at Half Maximum (FWHM) Gaussian kernel.

638 During first-level analysis, we included six task (block) regressors that were
639 composed of 3 task conditions by positive and negative valence. The task regressors were
640 convolved with a canonical hemodynamic response function. We also included six
641 additional regressors for motion parameters. A high-pass frequency filter (128s) and
642 autocorrelation corrections (using restricted maximum likelihood and an autoregressive
643 model) were used in model estimation.

644

645 ***Validation dataset***

646 We conducted a non-systematic literature review to identify recent fMRI studies of
647 comparable mentalizing tasks with at least three conditions (Self, Other, and non-
648 mentalizing Control condition) and in which participants rated whether trait adjectives or
649 statements. We emailed the authors of seven studies that we identified and received
650 positive responses from four of them. Data from these four studies were included in the
651 current study as validation datasets. In addition, to test the signatures in a different type
652 of task, we included an unpublished dataset by one of the co-authors (JAH) as the
653 extension dataset (Study 6). Please refer to the original studies (see Supplementary Table
654 1) for details of image acquisition, preprocessing, and first-level analysis.

655 The authors provided the person-level contrast images of three conditions (versus
656 implicit baseline). Voxel weights were normalized using L2-norm. These contrast images
657 were resampled onto the same image space as the training dataset using linear
658 resampling. For Study 3, which included two different Other-conditions (similar and
659 dissimilar classmates), we averaged the contrast images across these two conditions,
660 resulting in a single Other-condition, as in the training and the other validation datasets
661 (analyzing them separately did not alter the results).

662

663 **Data Analysis**

664 ***Training and cross-validation***

665 Using 10-fold cross-validation, we trained three support-vector-machine (SVM)
666 classifiers in Study 1 that discriminate each condition from the other two conditions: The
667 Self-RS was trained to separate the Self-condition from Other and Control conditions, the
668 Other-RS to separate the Other-condition from Self and Control conditions, and the MS
669 was trained to separate both mentalizing conditions (Self and Other) from the Control
670 condition. To reduce the possibility that classifiers opportunistically used non-mentalizing
671 related processes (e.g., visual information), we applied a mask in the training dataset that
672 includes key social-cognition regions (see Fig. 1b). This mask was computed as the union
673 of six term-based meta-analytic maps (association and uniformity maps for “mentalizing”,
674 “self-referential”, and “social”, downloaded from NeuroSynth [Yarkoni et al., 2011;
675 <https://neurosynth.org>] on 06/06/2024).

676 SVM were chosen based on their high performance for binary linear classification
677 problems in high dimensional data. The algorithm fits a hyperplane that classifies true and
678 false classes by assigning weights for each feature (i.e., each voxel). The fitting is
679 performed for each of 10 folds, in which the data of 90% of participants are used for fitting
680 the classifier and the resulting classifier is tested on the remaining 10% hold out
681 participants’ data, allowing to assess its classification performance in independent hold-
682 out data. Because using a one versus the rest approach in SVM (e.g., Self versus Other
683 and Control, see Fig. 1c) may add bias into the model by favoring the majority class, we
684 fitted weighted SVM models with a ridge amount of .5. To avoid overfitting, the SVMs were
685 otherwise trained using default parameters (regularization parameter $C = 1$). The cross-
686 validated distance from the hyperplane of hold-out images was used to calculate the
687 receiver operating characteristic (ROC) curves and the accuracy for each classification.
688 Each SVM results in a weight map with one value per voxel. These voxel weights are
689 effectively the predictive weights of the true class, yielding brain signatures of mentalizing
690 that can be applied to other brain images to obtain a single pattern expression score per
691 image.

692 We used bootstrapping to illustrate the regions that most significantly contribute to
693 the classification. Statistical weight maps were calculated using 5000 bootstrapped
694 samples that yielded two-tailed, uncorrected p -values for each voxel. These maps were
695 then thresholded using FDR-correction at $q < .05$ with a minimum cluster size of 10 voxels.
696 Note that corrected maps were only used to illustrate the most important contributing
697 regions of each signature. Unthresholded weight maps are used for classification and
698 hence constitute the brain signatures used for all further steps of the present analyses and

699 which can be used in future tests in other studies. The classification accuracies in training
700 dataset were tested using two-alternative forced-choice predictions and binomial tests.

701

702 ***Validation in independent datasets***

703 The three mentalizing signatures were applied to the validation datasets by
704 computing the pattern similarity values as the matrix dot product between mentalizing
705 signatures and the person-level (1st level) contrast images of each study, yielding one
706 scalar value per condition and participant and signature. The predictions followed a binary
707 forced-choice principle using paired observations. Lastly, the signatures' classification
708 accuracies were assessed using Receiver Operating Characteristic (ROC) analysis and
709 binomial tests using a two-sided significance threshold of $p < .05$.

710 ***Group comparisons of pattern expression values.*** To quantify how well
711 signatures discriminate Self versus Other conditions and to easily compare self/other
712 separation across groups, we calculated a 'true minus false class score' for each signature
713 by subtracting the pattern expression values of the false condition images from the true
714 condition images (i.e., responses of the Self-RS for the Self minus Other condition,
715 responses of the Other-RS for Other minus Self condition, responses of the MS for the
716 mean of Self and Other conditions minus Control condition). For statistical comparisons
717 between participants with schizophrenia and healthy controls, we ran linear mixed effects
718 models for each of the signatures, with self-other discrimination ('true minus false class
719 score') as the dependent variable. The group constituted the fixed effect variable (healthy
720 controls [$n=48$] versus schizophrenia sample [$n=40$]). Because the data came from two
721 different studies (Study 4 and Study 5), study was included as a random effect variable.
722 Therefore, the model equation was "*true_minus_false_class_score* ~ *group* + (1|*study*)".
723 To test whether the bipolar sample ($n=18$) differed from healthy adults ($n=15$) in Study 5
724 regarding their pattern expression values, we performed independent samples t-test for
725 each of the signatures consecutively. The dependent variable was the "true minus false
726 class score" mentioned above.

727 ***Associations with age.*** We tested whether the adolescents' age was associated
728 with self-other discrimination using linear mixed effects models. We included age as the
729 fixed effect, and Study (Study 2 and Study 3) as a random effect. The dependent variable
730 was the difference of the true minus false classes for each of the signature, as detailed
731 above. Therefore, the model formula was "*true_minus_false_class_score* ~ *ages* +
732 (1|*study*)".

733

734 *Testing the signature in an extension task (social feedback, Study 6)*

735 The signatures were applied to the subject-level contrast images for three
736 experimental conditions (feedback to the self, feedback to the partner, no feedback), by
737 computing their dot products. The resulting pattern expression values were subjected to
738 repeated-measures ANOVAs to test for differences across the three conditions of the
739 social-feedback task. Bonferroni-corrected t-tests were used for subsequent pairwise
740 comparisons.

741

742 *Region-of-Interest (ROI) Analyses*

743 To test the ability of local patterns to predict mentalizing and to separate self-
744 related versus other-related mentalizing, we trained and cross-validated region-of-interest
745 (ROI) classifiers using a parallel approach to the whole-brain classifiers. We downloaded
746 a term-based meta-analytic map for 'Mentalizing' from NeuroSynth on 14/09/2022 that
747 included 151 studies. We selected clusters that contained more than 200 voxels, resulting
748 in the following ten ROIs: mPFC, bilateral TPJ, bilateral anterior MTG, precuneus/PCC,
749 right SMA, and three clusters in the cerebellum. In each ROI, we trained and 10-fold cross-
750 validated four classifiers in the training dataset and validated them in the remaining
751 datasets, to (i) predict Self-mentalizing (versus the two remaining conditions), (ii) predict
752 Other-mentalizing (versus the two remaining conditions), (iii) predict mentalizing (Self- and
753 Other-mentalizing versus Control condition), and (iv) differentiate Self- versus Other-
754 processing. We tested these ROI classifiers in the validation datasets using the same
755 parameters and analytic approach as outlined above in the main whole-brain analyses.

756

757 *General statistical approach*

758 All data analysis was performed using Matlab® R2022b software and the Canlab
759 toolbox (<https://github.com/canlab>). Statistical inference used a significance threshold of
760 $p < 0.05$, unless otherwise noted.

761

762

References

- 763 Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010).
 764 Functional-Anatomic Fractionation of the Brain's Default Network. *Neuron*, 65(4),
 765 550–562. <https://doi.org/10.1016/j.neuron.2010.02.005>
- 766 Arioli, M., Cattaneo, Z., Parimbelli, S., & Canessa, N. (2023). Relational vs
 767 representational social cognitive processing: A coordinate-based meta-analysis
 768 of neuroimaging data. *Social Cognitive and Affective Neuroscience*, 18(1),
 769 nsad003. <https://doi.org/10.1093/scan/nsad003>
- 770 Arioli, M., Cattaneo, Z., Ricciardi, E., & Canessa, N. (2021). Overlapping and specific
 771 neural correlates for empathizing, affective mentalizing, and cognitive
 772 mentalizing: A coordinate-based meta-analytic study. *Human Brain Mapping*,
 773 42(14), 4777–4804. <https://doi.org/10.1002/hbm.25570>
- 774 Babo-Rebelo, M., & Tallon-Baudry, C. (2018). Interoceptive signals, brain dynamics, and
 775 subjectivity. In M. Tsakiris & H. De Preester (Eds.), *The Interoceptive Mind: From*
 776 *Homeostasis to Awareness* (pp. 46–62). Oxford University Press.
- 777 Beck, A. T., Baruch, E., Balter, J. M., Steer, R. A., & Warman, D. M. (2004). A new
 778 instrument for measuring insight: The Beck Cognitive Insight Scale.
 779 *Schizophrenia Research*, 68(2–3), 319–329. [https://doi.org/10.1016/S0920-](https://doi.org/10.1016/S0920-9964(03)00189-0)
 780 [9964\(03\)00189-0](https://doi.org/10.1016/S0920-9964(03)00189-0)
- 781 Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia:
 782 Meta-analysis. *Schizophrenia Research*, 109(1), 1–9.
 783 <https://doi.org/10.1016/j.schres.2008.12.020>
- 784 Brüne, M., & Brüne-Cohrs, U. (2006). Theory of mind—Evolution, ontogeny, brain
 785 mechanisms and psychopathology. *Neuroscience & Biobehavioral Reviews*,
 786 30(4), 437–455. <https://doi.org/10.1016/j.neubiorev.2005.08.001>
- 787 Chand, G. B., Singhal, P., Dwyer, D. B., Wen, J., Erus, G., Doshi, J., Srinivasan, D.,
 788 Mamourian, E., Varol, E., Sotiras, A., Hwang, G., Dazzan, P., Kahn, R. S.,
 789 Schnack, H. G., Zanetti, M. V., Meisenzahl, E., Busatto, G. F., Crespo-Facorro,
 790 B., Pantelis, C., ... Davatzikos, C. (2022). Schizophrenia Imaging Signatures and
 791 Their Associations With Cognition, Psychopathology, and Genetics in the
 792 General Population. *American Journal of Psychiatry*, 179(9), 650–660.
 793 <https://doi.org/10.1176/appi.ajp.21070686>
- 794 Chavez, R. S., Heatherton, T. F., & Wagner, D. D. (2017). Neural Population Decoding
 795 Reveals the Intrinsic Positivity of the Self. *Cerebral Cortex* 27(11), 5222–5229.
 796 <https://doi.org/10.1093/cercor/bhw302>
- 797 Corradi-Dell'Acqua, C., Hofstetter, C., & Vuilleumier, P. (2014). Cognitive and affective
 798 theory of mind share the same local patterns of activity in posterior temporal but
 799 not medial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, 9(8),
 800 1175–1184. <https://doi.org/10.1093/scan/nst097>
- 801 Courtney, A. L., & Meyer, M. L. (2020). Self-Other Representation in the Social Brain
 802 Reflects Social Connection. *The Journal of Neuroscience*, 40(29), 5616–5627.
 803 <https://doi.org/10.1523/JNEUROSCI.2826-19.2020>
- 804 Crone, E. A., & Fuligni, A. J. (2020). Self and Others in Adolescence. *Annual Review of*
 805 *Psychology*, 71(1), 447–469. [https://doi.org/10.1146/annurev-psych-010419-](https://doi.org/10.1146/annurev-psych-010419-050937)
 806 [050937](https://doi.org/10.1146/annurev-psych-010419-050937)
- 807 Dabiri, M., Dehghani Firouzabadi, F., Yang, K., Barker, P. B., Lee, R. R., & Yousem, D.
 808 M. (2022). Neuroimaging in schizophrenia: A review article. *Frontiers in*
 809 *Neuroscience*, 16.
 810 <https://www.frontiersin.org/articles/10.3389/fnins.2022.1042814>

- 811 Debbané, M., Badoud, D., Sander, D., Eliez, S., Luyten, P., & Vrtička, P. (2017). Brain
812 activity underlying negative self- and other-perception in adolescents: The role of
813 attachment-derived self-representations. *Cognitive, Affective and Behavioral*
814 *Neuroscience*, 17(3), 554–576. [https://doi.org/10.3758/S13415-017-0497-](https://doi.org/10.3758/S13415-017-0497-9/FIGURES/6)
815 [9/FIGURES/6](https://doi.org/10.3758/S13415-017-0497-9/FIGURES/6)
- 816 Debbané, M., Salaminios, G., Luyten, P., Badoud, D., Armando, M., Solida Tozzi, A.,
817 Fonagy, P., & Brent, B. K. (2016). Attachment, Neurobiology, and Mentalizing
818 along the Psychosis Continuum. *Frontiers in Human Neuroscience*, 10.
819 <https://doi.org/10.3389/fnhum.2016.00406>
- 820 Defendini, A., & Jenkins, A. C. (2023). Dissociating neural sensitivity to target identity
821 and mental state content type during inferences about other minds. *Social*
822 *Neuroscience*, 18(2), 103–121. <https://doi.org/10.1080/17470919.2023.2208879>
- 823 Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of
824 functional neuroimaging studies of self- and other judgments reveals a spatial
825 gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive*
826 *Neuroscience*, 24(8), 1742–1752. https://doi.org/10.1162/JOCN_A_00233
- 827 Fehlbauer, L. V., Borbás, R., Paul, K., Eickhoff, S. B., & Raschle, N. M. (2022). Early and
828 late neural correlates of mentalizing: ALE meta-analyses in adults, children and
829 adolescents. *Social Cognitive and Affective Neuroscience*, 17(4), 351–366.
830 <https://doi.org/10.1093/scan/nsab105>
- 831 First, M.B. (2015). Structured Clinical Interview for the DSM (SCID). In The Encyclopedia
832 of Clinical Psychology (eds R.L. Cautin and S.O. Lilienfeld). <https://doi.org/10.1002/9781118625392.wbecp351>
- 834 Fossati, P., Hevenor, S. J., Graham, S. J., Grady, C., Keightley, M. L., Craik, F., &
835 Mayberg, H. (2003). In Search of the Emotional Self: An fMRI Study Using
836 Positive and Negative Emotional Words. *American Journal of Psychiatry*,
837 160(11), 1938–1945. <https://doi.org/10.1176/appi.ajp.160.11.1938>
- 838 Frith, C. D., & Frith, U. (2006). The Neural Basis of Mentalizing. *Neuron*, 50(4), 531–534.
839 <https://doi.org/10.1016/j.neuron.2006.05.001>
- 840 Fuentes-Claramonte, P., Martín-Subero, M., Salgado-Pineda, P., Alonso-Lana, S.,
841 Moreno-Alcázar, A., Argila-Plaza, I., Santo-Angles, A., Albajes-Eizagirre, A.,
842 Anguera-Camós, M., Capdevila, A., Sarró, S., McKenna, P. J., Pomarol-Clotet,
843 E., & Salvador, R. (2019). Shared and differential default-mode related patterns
844 of activity in an autobiographical, a self-referential and an attentional task. *PLOS*
845 *ONE*, 14(1), e0209376. <https://doi.org/10.1371/journal.pone.0209376>
- 846 Fuentes-Claramonte, P., Martín-Subero, M., Salgado-Pineda, P., Santo-Angles, A.,
847 Argila-Plaza, I., Salavert, J., Arévalo, A., Bosque, C., Sarri, C., Guerrero-
848 Pedraza, A., Capdevila, A., Sarró, S., McKenna, P. J., Pomarol-Clotet, E., &
849 Salvador, R. (2020). Brain imaging correlates of self- and other-reflection in
850 schizophrenia. *NeuroImage: Clinical*, 25, 102134.
851 <https://doi.org/10.1016/j.nicl.2019.102134>
- 852 Gabrieli, J. D. E., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a
853 Humanitarian and Pragmatic Contribution from Human Cognitive Neuroscience.
854 *Neuron*, 85(1), 11–26. <https://doi.org/10.1016/j.neuron.2014.10.047>
- 855 Garfinkel, S. N., Nagai, Y., Seth, A. K., & Critchley, H. D. (2013). Neuroimaging Studies
856 of Interoception and Self-Awareness. In A. E. Cavanna, A. Nani, H. Blumenfeld,
857 & S. Laureys (Eds.), *Neuroimaging of Consciousness* (pp. 207–224). Springer.
858 https://doi.org/10.1007/978-3-642-37580-4_11
- 859 Gray, K., Jenkins, A. C., Heberlein, A. S., & Wegner, D. M. (2011). Distortions of mind
860 perception in psychopathology. *Proceedings of the National Academy of*
861 *Sciences*, 108(2), 477–479. <https://doi.org/10.1073/pnas.1015493108>

- 862 Heatherton, T. F., Wyland, C. L., Macrae, C. N., Demos, K. E., Denny, B. T., & Kelley,
863 W. M. (2006). Medial prefrontal activity differentiates self from close others.
864 *Social Cognitive and Affective Neuroscience*, 1(1), 18–25.
865 <https://doi.org/10.1093/scan/nsl001>
- 866 Johnson, B. N., Kivity, Y., Rosenstein, L. K., LeBreton, J. M., & Levy, K. N. (2022). The
867 association between mentalizing and psychopathology: A meta-analysis of the
868 reading the mind in the eyes task across psychiatric disorders. *Clinical*
869 *Psychology: Science and Practice*, 29(4), 423–439.
870 <https://doi.org/10.1037/cps0000105>
- 871 Kemp, R., & David, A. (1997). Insight and compliance. In B. Blackwell (Ed.), *Treatment*
872 *compliance and the therapeutic alliance* (pp. 61–84). Harwood Academic
873 Publishers.
- 874 Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F.
875 (2002). Finding the Self? An Event-Related fMRI Study. *Journal of Cognitive*
876 *Neuroscience*, 14(5), 785–794. <https://doi.org/10.1162/08989290260138672>
- 877 Kim, H. J., Lux, B. K., Lee, E., Finn, E. S., & Woo, C.-W. (2024). Brain decoding of
878 spontaneous thought: Predictive modeling of self-relevance and valence using
879 personal narratives. *Proceedings of the National Academy of Sciences*, 121(14),
880 e2401959121. <https://doi.org/10.1073/pnas.2401959121>
- 881 Kim, J., Andrews-Hanna, J. R., Eisenbarth, H., Lux, B. K., Kim, H. J., Lee, E., Lindquist,
882 M. A., Losin, E. A. R., Wager, T. D., & Woo, C.-W. (2023). A dorsomedial
883 prefrontal cortex-based dynamic functional connectivity model of rumination.
884 *Nature Communications*, 14(1), 3540. [https://doi.org/10.1038/s41467-023-39142-](https://doi.org/10.1038/s41467-023-39142-9)
885 [9](https://doi.org/10.1038/s41467-023-39142-9)
- 886 Kim Lux, B., Andrews-Hanna, J. R., Han, J., Lee, E., & Woo, C.-W. (2022). When self
887 comes to a wandering mind: Brain representations and dynamics of self-
888 generated concepts in spontaneous thought. *Science Advances*, 8(35),
889 eabn8616. <https://doi.org/10.1126/sciadv.abn8616>
- 890 Koban, L., Gianaros, P. J., Kober, H., & Wager, T. D. (2021). The self in context: Brain
891 systems linking mental and physical health. *Nature Reviews Neuroscience*.
892 <https://doi.org/10.1038/s41583-021-00446-8>
- 893 Koban, L., Pichon, S., & Vuilleumier, P. (2014). Responses of medial and ventrolateral
894 prefrontal cortex to interpersonal conflict for resources. *Social Cognitive and*
895 *Affective Neuroscience*, 9(5), 561–569. <https://doi.org/10.1093/scan/nst020>
- 896 Koban, L., Wager, T. D., & Kober, H. (2023). A neuromarker for drug and food craving
897 distinguishes drug users from non-users. *Nature Neuroscience*, 26(2), Article 2.
898 <https://doi.org/10.1038/s41593-022-01228-w>
- 899 Kragel, P. A., Koban, L., Barrett, L. F., & Wager, T. D. (2018). Representation, Pattern
900 Information, and Brain Signatures: From Neurons to Neuroimaging. *Neuron*,
901 99(2), 257–273. <https://doi.org/10.1016/j.neuron.2018.06.009>
- 902 Luyten, P., Campbell, C., Allison, E., & Fonagy, P. (2020). The Mentalizing Approach to
903 Psychopathology: State of the Art and Future Directions. *Annual Review of*
904 *Clinical Psychology*, 16(1), 297–325. [https://doi.org/10.1146/annurev-clinpsy-](https://doi.org/10.1146/annurev-clinpsy-071919-015355)
905 [071919-015355](https://doi.org/10.1146/annurev-clinpsy-071919-015355)
- 906 Ma, S., Maresh, E. L., Coppola, A. M., Richard, K. E., Koban, L., Sbarra, D., & Andrews-
907 Hanna, J. R. (2024). *When Empathy Gets Tough: Neural Responses to*
908 *Overcoming the Self in a Novel Paradigm Predict Everyday Prosocial Behavior*.
909 PsyArXiv. doi.org/10.31234/osf.io/qcj45
- 910 Maresh, E. L., & Andrews-Hanna, J. R. (2021). Putting the “Me” in “Mentalizing”: Multiple
911 Constructs Describing Self Versus Other During Mentalizing and Implications for
912 Social Anxiety Disorder. In M. Gilead & K. N. Ochsner (Eds.), *The Neural Basis*

- 913 *of Mentalizing* (pp. 629–658). Springer International Publishing.
 914 https://doi.org/10.1007/978-3-030-51890-5_33
- 915 Molapour, T., Hagan, C. C., Silston, B., Wu, H., Ramstead, M., Friston, K., & Mobbs, D.
 916 (2021). Seven computations of the social brain. *Social Cognitive and Affective*
 917 *Neuroscience*, 16(8), 745–760. <https://doi.org/10.1093/scan/nsab024>
- 918 Moore, C., & Frye, D. (1991). The Acquisition and Utility of Theories of Mind. In D. Frye
 919 & C. Moore (Eds.), *Children's Theories of Mind: Mental States and Social*
 920 *Understanding* (pp. 1–14). Lawrence Erlbaum.
- 921 Murray, R. J., Debbané, M., Fox, P. T., Bzdok, D., & Eickhoff, S. B. (2014). Functional
 922 connectivity mapping of regions associated with self- and other-processing.
 923 *Human Brain Mapping*, 36(4), 1304–1324. <https://doi.org/10.1002/hbm.22703>
- 924 Northoff, G., Heinzl, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J.
 925 (2006). Self-referential processing in our brain—A meta-analysis of imaging
 926 studies on the self. *NeuroImage*, 31(1), 440–457.
 927 <https://doi.org/10.1016/j.neuroimage.2005.12.002>
- 928 Northoff, G., & Hayes, D. J. (2011). Is Our Self Nothing but Reward? *Biological*
 929 *Psychiatry*, 69(11), 1019–1025. <https://doi.org/10.1016/j.biopsych.2010.12.014>
- 930 Oosterwijk, S., Snoek, L., Rotteveel, M., Barrett, L. F., & Scholte, H. S. (2017). Shared
 931 states: Using MVPA to test neural overlap between self-focused emotion imagery
 932 and other-focused emotion understanding. *Social Cognitive and Affective*
 933 *Neuroscience*, 12(7), 1025–1035. <https://doi.org/10.1093/scan/nsx037>
- 934 Parelman, J. M., Doré, B. P., Cooper, N., O'Donnell, M. B., Chan, H.-Y., & Falk, E. B.
 935 (2021). Overlapping Functional Representations of Self- and Other-Related
 936 Thought are Separable Through Multivoxel Pattern Classification. *Cerebral*
 937 *Cortex*. <https://doi.org/10.1093/CERCOR/BHAB272>
- 938 Peer, M., Salomon, R., Goldberg, I., Blanke, O., & Arzy, S. (2015). Brain system for
 939 mental orientation in space, time, and person. *Proceedings of the National*
 940 *Academy of Sciences*, 112(35), 11072–11077.
 941 <https://doi.org/10.1073/pnas.1504242112>
- 942 Potvin, S., Gamache, L., & Lungu, O. (2019). A Functional Neuroimaging Meta-Analysis
 943 of Self-Related Processing in Schizophrenia. *Frontiers in Neurology*, 10.
 944 <https://doi.org/10.3389/fneur.2019.00990>
- 945 Qin, P., Wang, M., & Northoff, G. (2020). Linking bodily, environmental and mental
 946 states in the self—A three-level model based on a meta-analysis. *Neuroscience*
 947 *& Biobehavioral Reviews*, 115, 77–95.
 948 <https://doi.org/10.1016/j.neubiorev.2020.05.004>
- 949 Quesque, F., Apperly, I., Baillargeon, R., Baron-Cohen, S., Becchio, C., Bekkering, H.,
 950 Bernstein, D., Bertoux, M., Bird, G., Bukowski, H., Burgmer, P., Carruthers, P.,
 951 Catmur, C., Dziobek, I., Epley, N., Erle, T. M., Frith, C., Frith, U., Galang, C. M.,
 952 ... Brass, M. (2024). Defining key concepts for mental state attribution.
 953 *Communications Psychology*, 2(1), 1–5. [https://doi.org/10.1038/s44271-024-](https://doi.org/10.1038/s44271-024-00077-6)
 954 [00077-6](https://doi.org/10.1038/s44271-024-00077-6)
- 955 Rosenberg, M. D., Finn, E. S., Scheinost, D., Papademetris, X., Shen, X., Constable, R.
 956 T., & Chun, M. M. (2016). A neuromarker of sustained attention from whole-brain
 957 functional connectivity. *Nature Neuroscience*, 19(1), 165–171.
 958 <https://doi.org/10.1038/nn.4179>
- 959 Rosenberg, M. D., Casey, B. J., & Holmes, A. J. (2018). Prediction complements
 960 explanation in understanding the developing brain. *Nature Communications*, 9(1),
 961 589. <https://doi.org/10.1038/s41467-018-02887-9>
- 962 Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*,
 963 16(2), 235–239. <https://doi.org/10.1016/j.conb.2006.03.001>

- 964 Scheinost, D., Noble, S., Horien, C., Greene, A. S., Lake, E. Mr., Salehi, M., Gao, S.,
 965 Shen, X., O'Connor, D., Barron, D. S., Yip, S. W., Rosenberg, M. D., &
 966 Constable, R. T. (2019). Ten simple rules for predictive modeling of individual
 967 differences in neuroimaging. *NeuroImage*, 193, 35–45.
 968 <https://doi.org/10.1016/j.neuroimage.2019.02.057>
- 969 Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating
 970 theory of mind: A meta-analysis of functional brain imaging studies.
 971 *Neuroscience & Biobehavioral Reviews*, 42, 9–34.
 972 <https://doi.org/10.1016/j.neubiorev.2014.01.009>
- 973 Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet,
 974 J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A
 975 neuroimaging meta-analysis and integrative review of empathy and theory of
 976 mind. *Psychological Bulletin*, 147(3), 293–327.
 977 <https://doi.org/10.1037/bul0000303>
- 978 Sharp, C. (2006). Mentalizing Problems in Childhood Disorders. In *Handbook of*
 979 *Mentalization-Based Treatment* (pp. 101–121). John Wiley & Sons, Ltd.
 980 <https://doi.org/10.1002/9780470712986.ch4>
- 981 Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E.,
 982 Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International
 983 Neuropsychiatric Interview (M.I.N.I.): The development and validation of a
 984 structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of*
 985 *Clinical Psychiatry*, 59 Suppl 20, 22-33;quiz 34-57.
- 986 Sloover, M., van Est, L. A. C., Janssen, P. G. J., Hilbink, M., & van Ee, E. (2022). A
 987 meta-analysis of mentalizing in anxiety disorders, obsessive-compulsive and
 988 related disorders, and trauma and stressor related disorders. *Journal of Anxiety*
 989 *Disorders*, 92, 102641. <https://doi.org/10.1016/j.janxdis.2022.102641>
- 990 Sprong, M., Schothorst, P., Vos, E., Hox, J., & Engeland, H. V. (2007). Theory of mind in
 991 schizophrenia: Meta-analysis. *The British Journal of Psychiatry*, 191(1), 5–13.
 992 <https://doi.org/10.1192/bjp.bp.107.035899>
- 993 Tamir, D. I., Bricker, A. B., Dodell-Feder, D., & Mitchell, J. P. (2016). Reading fiction and
 994 reading minds: The role of simulation in the default network. *Social Cognitive and*
 995 *Affective Neuroscience*, 11(2), 215–224. <https://doi.org/10.1093/scan/nsv114>
- 996 Tamir, D. I., & Mitchell, J. P. (2012). Disclosing information about the self is intrinsically
 997 rewarding. *Proceedings of the National Academy of Sciences*, 109(21), 8038–
 998 8043. <https://doi.org/10.1073/pnas.1202129109>
- 999 Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in*
 1000 *Cognitive Sciences*, 22(3), 201–212. <https://doi.org/10.1016/j.tics.2017.12.005>
- 1001 Tan, K. M., Daitch, A. L., Pinheiro-Chagas, P., Fox, K. C. R., Parvizi, J., & Lieberman, M.
 1002 D. (2022). Electrographic evidence of a common neurocognitive sequence
 1003 for mentalizing about the self and others. *Nature Communications*, 13(1), 1919.
 1004 <https://doi.org/10.1038/s41467-022-29510-2>
- 1005 van Buuren, M., Walsh, R. J., Sijtsma, H., Hollarek, M., Lee, N. C., Bos, P. A., &
 1006 Krabbendam, L. (2020). Neural correlates of self- and other-referential
 1007 processing in young adolescents and the effects of testosterone and peer
 1008 similarity. *NeuroImage*, 219, 117060.
 1009 <https://doi.org/10.1016/j.neuroimage.2020.117060>
- 1010 Van Der Meer, L., Costafreda, S., Aleman, A., & David, A. S. (2010). Self-reflection and
 1011 the brain: A theoretical review and meta-analysis of neuroimaging studies with
 1012 implications for schizophrenia. *Neuroscience & Biobehavioral Reviews*, 34(6),
 1013 935–946. <https://doi.org/10.1016/j.neubiorev.2009.12.004>

- 1014 Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by
 1015 mirror and mentalizing systems: A meta-analysis. *NeuroImage*, 48(3), 564–584.
 1016 <https://doi.org/10.1016/j.neuroimage.2009.06.009>
- 1017 Van Overwalle, F., & Vandekerckhove, M. (2013). Implicit and explicit social mentalizing:
 1018 Dual processes driven by a shared neural network. *Frontiers in Human*
 1019 *Neuroscience*, 7. <https://doi.org/10.3389/fnhum.2013.00560>
- 1020 Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An
 1021 fMRI-Based Neurologic Signature of Physical Pain. *New England Journal of*
 1022 *Medicine*, 368(15), 1388–1397. <https://doi.org/10.1056/NEJMoa1204471>
- 1023 Wagner, D. D., Chavez, R. S., & Broom, T. W. (2019). Decoding the neural
 1024 representation of self and person knowledge with multivariate pattern analysis
 1025 and data-driven approaches. *Wiley Interdisciplinary Reviews: Cognitive Science*,
 1026 10(1). <https://doi.org/10.1002/WCS.1482>
- 1027 Wang, Y., Metoki, A., Xia, Y., Zang, Y., He, Y., & Olson, I. R. (2021). A large-scale
 1028 structural and functional connectome of social mentalizing. *NeuroImage*, 236,
 1029 118115. <https://doi.org/10.1016/j.neuroimage.2021.118115>
- 1030 Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University
 1031 Press.
- 1032 Woo, C. W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better
 1033 biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience*,
 1034 20(3), 365. <https://doi.org/10.1038/NN.4478>
- 1035 Yang, D. Y.-J., Rosenblau, G., Keifer, C., & Pelphrey, K. A. (2015). An integrative neural
 1036 model of social perception, action observation, and theory of mind. *Neuroscience*
 1037 *& Biobehavioral Reviews*, 51, 263–275.
 1038 <https://doi.org/10.1016/j.neubiorev.2015.01.020>
- 1039 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011).
 1040 Large-scale automated synthesis of human functional neuroimaging data. *Nature*
 1041 *Methods*, 8(8), 665–670. <https://doi.org/10.1038/nmeth.1635>
- 1042 Zhang, L., Opmeer, E. M., Ruhé, H. G., Aleman, A., & van der Meer, L. (2015). Brain
 1043 activation during self- and other-reflection in bipolar disorder with a history of
 1044 psychosis: Comparison to schizophrenia. *NeuroImage: Clinical*, 8, 202–209.
 1045 <https://doi.org/10.1016/j.nicl.2015.04.010>

Figures

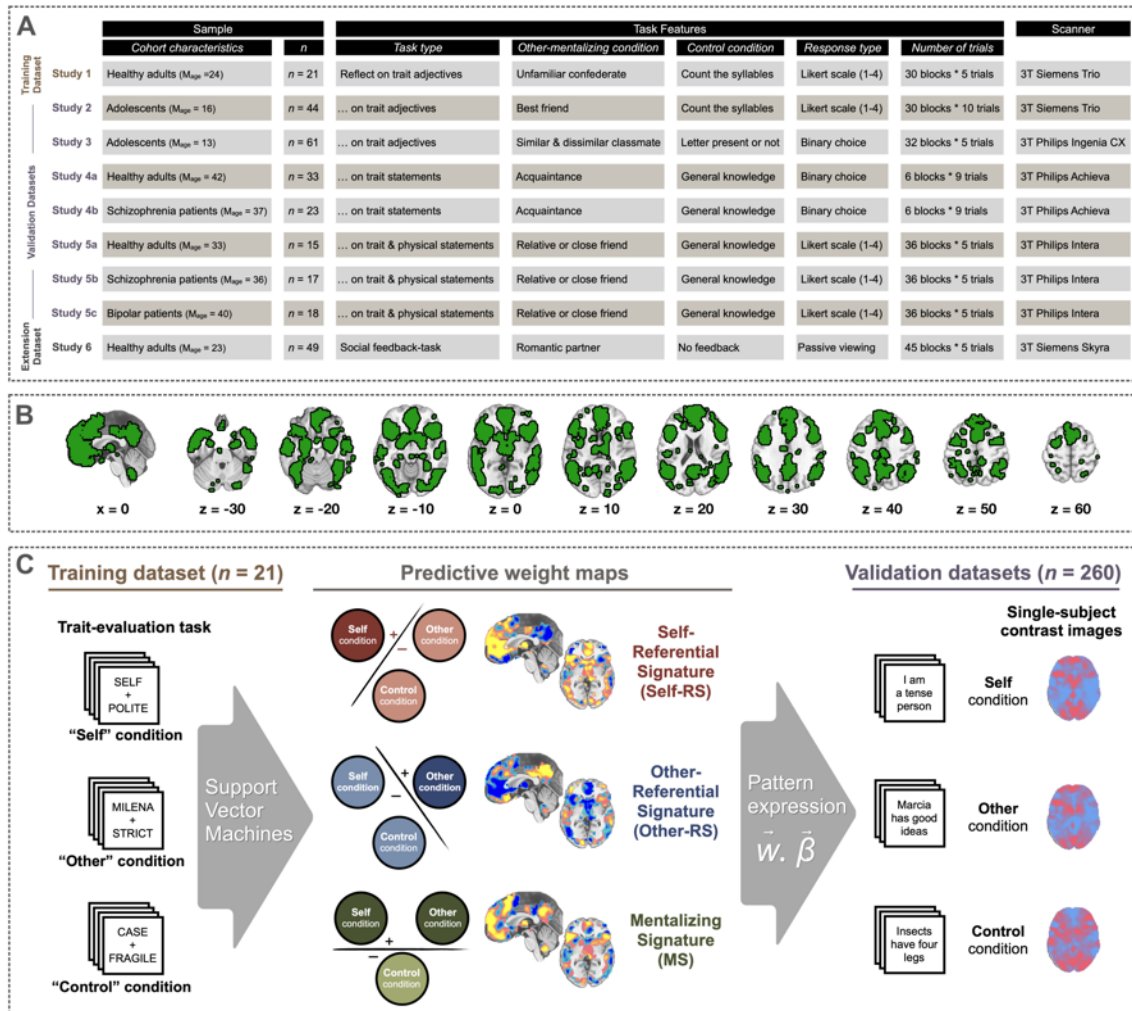


Figure 1. Study design and analytic approach. A) The present study included data from six independent studies (nine cohorts, total N = 281). All studies used a task that included a Self-condition, an Other-condition), and a non-social Control condition. B) Display of the inclusive mask of brain areas related to social cognition. C) The training dataset (Study 1) included contrast images of $n = 21$ participants performing a trait-evaluation task with three conditions (Self, Other, Control). We used 10-fold cross-validated linear SVM to train three mentalizing signatures. The Self-Referential Signature (Self-RS) was trained to predict the Self condition versus the two remaining conditions. The Other-Referential Signature (Other-RS) was trained to predict the Other condition versus two remaining conditions. The mentalizing signature (MS) was trained to separate the Self and Other conditions from Control condition. For the validation in independent datasets, we applied the signatures to the single-subject contrast images from Studies 2-6, by computing the dot product between the weight maps and the contrast images.

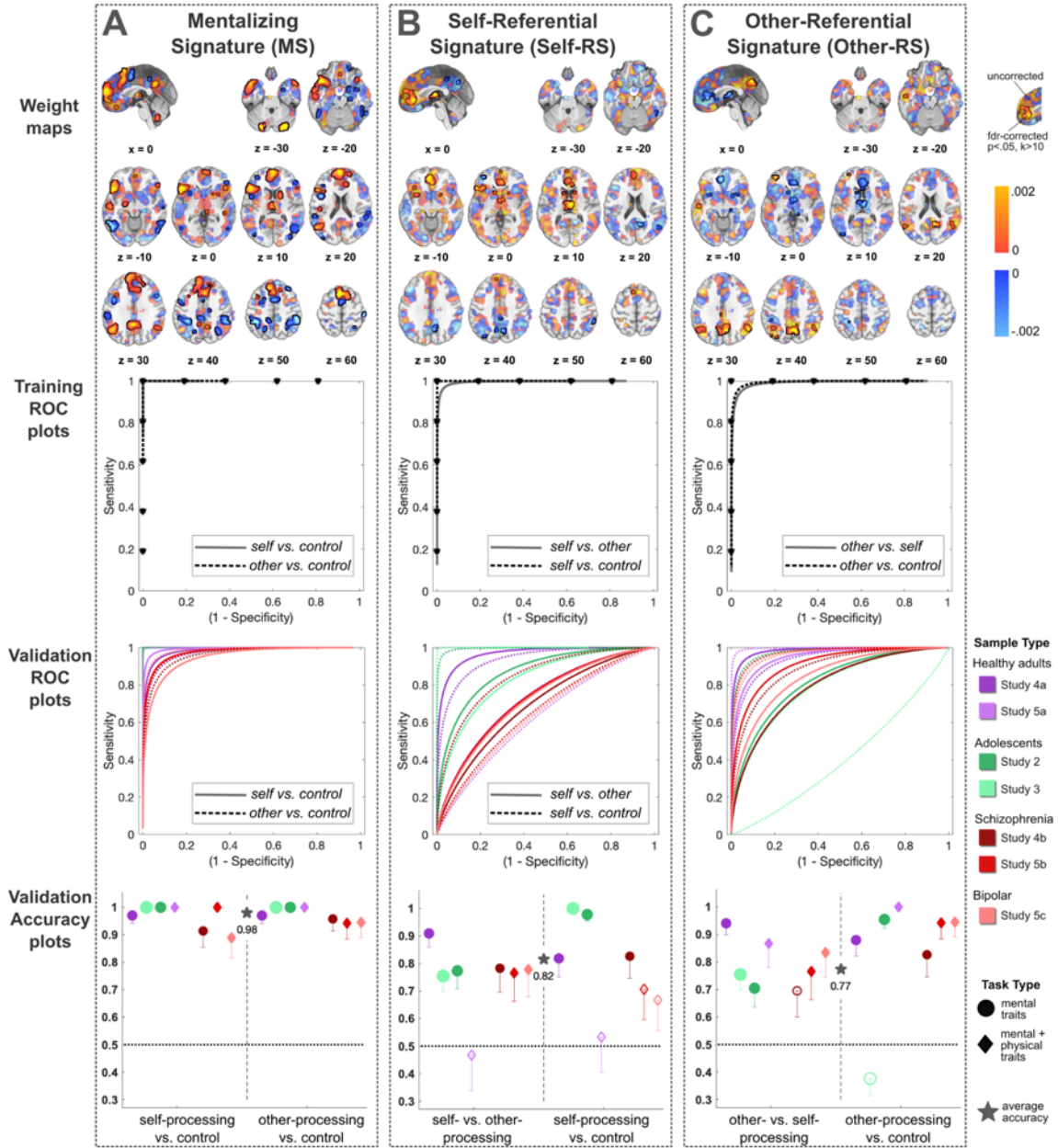


Figure 2. Model weights, training and validation results of the mentalizing signatures. On the top row, weight maps illustrate the positive and negative weights of a) the Mentalizing Signature (MS), b) the Self-Referential Signature (Self-RS), and c) the Other-Referential Signature (Other-RS). Voxels significant at an FDR-corrected threshold ($q < .05$, minimum cluster size of $k = 10$) are highlighted by black outlines. The two middle rows depict the receiver operation characteristics (ROC) plots from the i) cross-validated training and ii) validation datasets. The last row shows the accuracies of all three mentalizing signatures in each validation dataset separately, with color representing the study, and shape illustrating the type of sample.

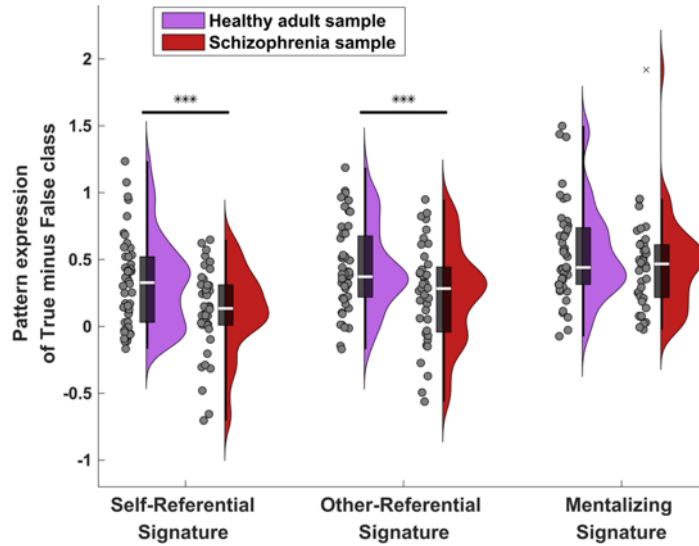


Figure 3. Self/other discrimination based on classifier responses in healthy adults versus individuals with schizophrenia. Self/other discrimination is measured as the pattern expression of true minus false class, separately for each signature and healthy adults ($n=48$, in red) versus individuals with schizophrenia ($n=40$, in red). Dots indicate values per person, boxplots mark the median, lower, and upper quartiles. Whiskers extend from the minimum to maximum data points excluding outliers. Linear mixed effects are used to test the group differences controlling for different cohorts. Self-RS: $\beta = .21$, $STE = .07$, $CI = [.07, .35]$, $p = .004$). Other-RS: $\beta = .19$, $STE = .07$, $CI = [.04, .33]$, $p = .01$.

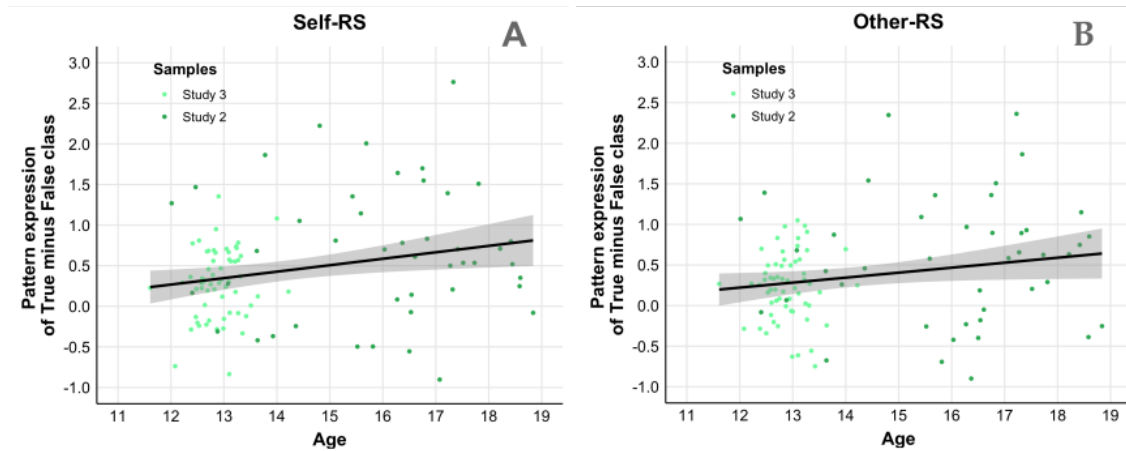


Figure 4. Association between self/other discrimination and age of adolescents ($n = 105$). Older age was significantly associated with better differentiation of self- and other-related mentalizing, controlling for cohort (Self-RS [$\beta = .08$, $STE = .03$, $CI = .012$ to $.018$, $p = .01$]; Other-RS [$\beta = .06$, $STE = .03$, $CI = .001$ to $.12$, $p = .047$]). Black lines display the linear regression fits, shaded areas the %95 confidence intervals.

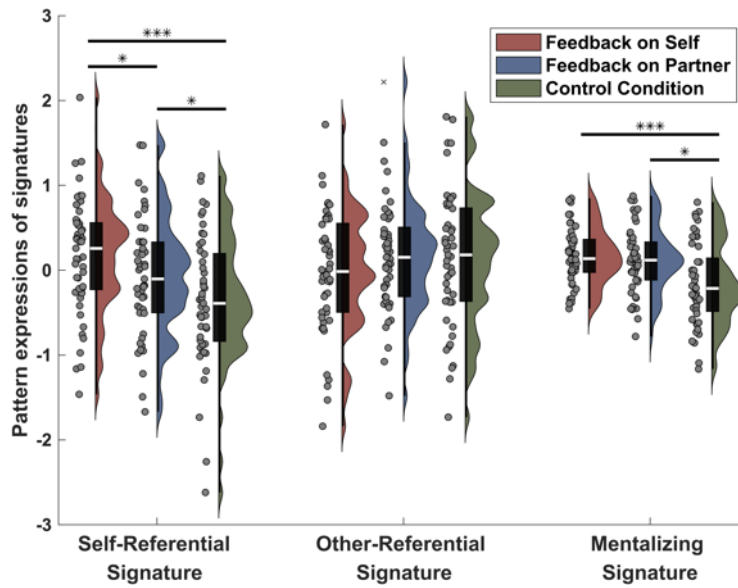


Figure 5. Signature responses in an independent social feedback task (n=49). Participants received feedback for themselves, for their partners, or no feedback (Control condition) and we computed pattern expression scores of each signature on each of these three conditions. Asterisks indicate significant Bonferroni-corrected pairwise comparisons following an initial repeated measures ANOVA test within each signature. * $p < .05$; *** $p < .001$.

BRAIN SIGNATURES OF MENTALIZING

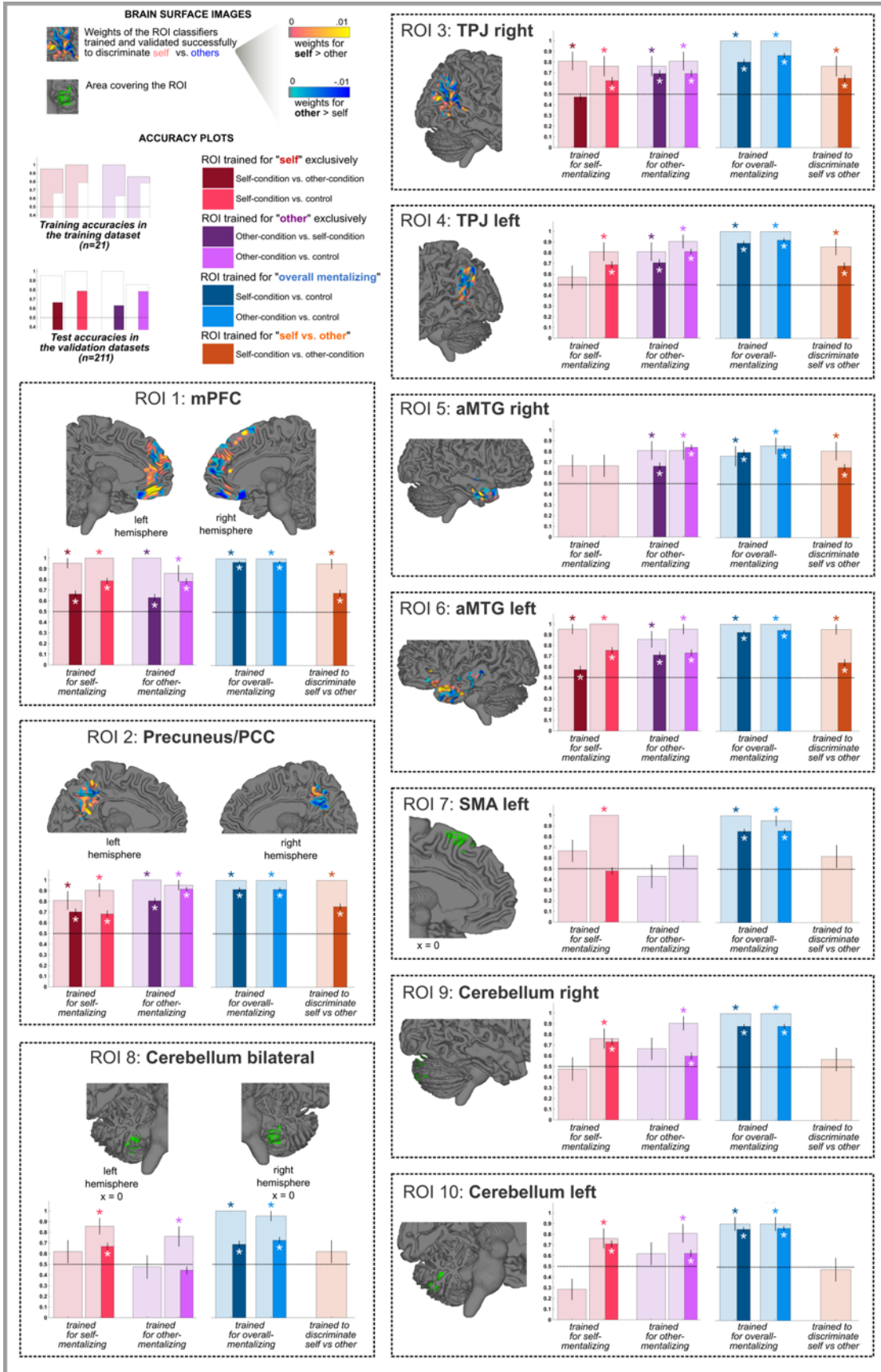


Figure 6. Results of the ROI analysis. Training and validation results for ten ROIs derived from a term-based meta-analysis ('mentalizing', NeuroSynth, Yarkoni, et al. 2011): mPFC, precuneus/posterior cingulate cortex (PCC), temporoparietal junction (TPJ) right, TPJ left, anterior middle temporal gyrus (aMTG) right, aMTG left, supplementary motor area (SMA) left, cerebellum right, cerebellum left, and a bilateral cerebellum cluster. Four different classifiers were trained for each ROI: (i) to predict Self exclusively, (ii) to predict Other exclusively, (iii) to predict overall mentalizing, and (iv) to discriminate the Self versus Other. If an ROI classifier had significant accuracy in the cross-validated training dataset, it was further tested across the validation datasets. The cross-validated training accuracies are illustrated by the wider bars in the background. The validation accuracies are illustrated by the thinner bars at the front. Classification performances significantly above the chance level (50%) are marked with asterisks. The surface images illustrate the weights of the classifiers that were trained to discriminate Self versus Other. Orange-yellow colors show areas with positive weights towards Self, blueish colors show areas associated with positive weights towards Other. If a specific ROI did not yield significant training and independent classification accuracy for this task, then the area covering this ROI is displayed in green color.

Supplementary Materials

Brain neuromarkers predict self- and other-related mentalizing across adult, clinical, and developmental samples

Dorukhan Açıl, Jessica Andrews-Hanna, Marina Lopez-Sola, Mariët van Buuren, Lydia Krabbendam, Liwen Zhang, Lisette van der Meer, Paola Fuentes-Claramonte, Edith Pomarol-Clotet, Raymond Salvador, Martin Debbané, Pascal Vrticka, Patrik Vuilleumier, David A. Sbarra, Andrea M. Coppola, Lars O. White, Tor D. Wager, & Leonie Koban*

*Please address correspondence to:

Dr. Leonie Koban, CRNL, Institut des Épilepsies IDEE, 59 Boulevard Pinel, 69500 Bron, France;
email: leonie.koban@cnrs.fr

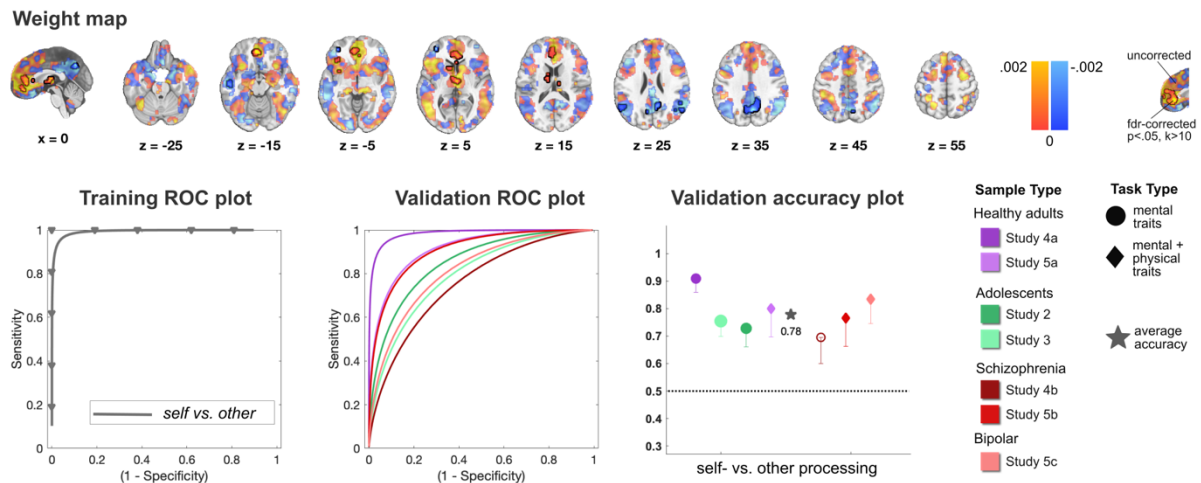


Figure S1. Model weights, training, and validation results of the Self-vs-Other Classifier. A classifier trained to separate self- versus other-related mentalizing (the *Self-vs-Other Signature*) showed 100% accuracy in cross-validated (out-of-sample) predictions in the training dataset (two-alternative forced-choice test, $p < .001$, Cohen's $d = 2.45$). The average prediction accuracy in the validation datasets was %78 (+/- %7.1 STE) and predicted the conditions above chance level in 6 out of 7 samples. Brain regions with significant positive voxel weights (associated with Self Condition) included ventromedial and ventrolateral prefrontal cortex, anterior cingulate cortex, left anterior insula, left inferior frontal gyrus, left caudate nucleus, left middle temporal cortex, and bilateral thalamus. Significant negative clusters (associated with Other Condition) were found in precuneus, bilateral temporoparietal junction/angular gyrus, right posterior cingulate cortex, left superior temporal sulcus, left middle frontal gyrus, and left supramarginal gyrus. The weight map (first row) illustrates the positive and negative weights, as well as the voxels significant at an FDR-corrected threshold ($q < .05$, minimum cluster size of $k = 10$). The second row depict the receiver operation characteristics (ROC) plots from the i) cross-validated training and ii) validation stages, and iii) the accuracies of Self vs Other Classifier in each validation dataset. Shapes in the validation accuracy plot encode different task types.

Table S1*Demographics*

Study	Population	n	Gender (females)	Age mean (SD)	Handedness (left)	Education (in years)	Place	References
Study 1	Healthy adults	21	10	23.7 (7.4)	2	NA	Switzerland	Koban et al., 2014
Study 2	Adolescents	44	23	16.0 (1.9)	0	NA	Switzerland	Debbane et al., 2017
Study 3	Adolescents	61	27	12.9 (0.4)	11	NA	The Netherlands	van Buuren et al., 2020
Study 4a	Healthy adults	33	14	41.7 (11.7)	0	12.7	Spain	Fuentes-Claramonte et al., 2019 & 2020
Study 4b	Schizophrenia patients	23	7	37.0 (8.1)	0	10.4		
Study 5a	Healthy adults	15	6	33.3 (11.3)	2	17.2	The Netherlands	Zhang et al., 2015
Study 5b	Schizophrenia patients	17	6	35.5 (9.7)	1	16.6		
Study 5c	Bipolar patients	18	9	40.3 (12.7)	0	17.0		
Study 6	Healthy adults	49	26	22.7 (3.9)	6	13.0	Arizona, USA	Ma et al., 2024
Overall		281	128	25.5 (12.9)	22			

Note. NA = not assessed. Links to the published studies are as follows: Study 2 (doi.org/10.3758/s13415-017-0497-9); Study 3 (doi.org/10.1016/j.neuroimage.2020.117060); Study 4, 2019 (doi.org/10.1371/journal.pone.0209376); Study 4, 2020 (doi.org/10.1016/j.nicl.2019.102134); Study 5 (doi.org/10.1016/j.nicl.2015.04.010); Study 6 (doi.org/10.31234/osf.io/qcj45)

Table S2*Significant clusters of the Self-RS*

Labeled Name	Atlas region (see note)	x	y	z	Voxel number	Max(z)
Cerebellum	Cblm Crus II R	26	-80	-38	54	0.00035
Ventral Anterior Cingulate Cortex (vACC) / ventromedial prefrontal cortex (vmPFC) (incl. dorsomedial and frontopolar prefrontal cortices)	Multiple regions	-2	44	0	933	0.00068
Anterior insula (AI)	AAIC L	-34	10	-10	33	0.00043
Caudate nucleus	Caudate Ca R	18	10	-6	13	0.00033
Inferior frontal gyrus (IFG)	45 L	-46	24	2	162	0.00057
Thalamus	Multiple regions	0	-12	10	311	0.00068
Caudate nucleus	Caudate Ca R	12	14	8	17	0.00031
Caudate nucleus	Caudate Ca L	-14	14	6	32	0.00035
Dorsolateral prefrontal cortex (dlPFC)	9m L	-8	48	38	21	0.00037
Superior frontal lobule/ dorsomedial prefrontal cortex (dmPFC)	SFL L	-8	22	56	46	0.00040
Temporal pole	TGd L	-34	14	-34	13	-0.00050
Inferior temporal gyrus/sulcus	TE2p R	46	-46	-22	27	-0.00045
Posterior temporal cortex	PH R	52	-54	-14	21	-0.00054
dlPFC	a9 46v L	-38	48	4	26	-0.00072
Posterior cingulate cortex (PCC)	POS2 R	6	-60	32	270	-0.00070
Inferior parietal cortex (IPC) / Temporoparietal junction (TPJ)	PGi R	52	-62	22	52	-0.00046
IPC / TPJ	PFm R	52	-50	24	20	-0.00035
IPC / TPJ	PGs L	-34	-76	40	14	-0.00044
PCC	31a L	-2	-42	42	20	-0.00054
IPC / TPJ	PFm R	52	-44	48	28	-0.00064

Note. Significant positive and negative weights contributing to the Self-Referential Signature (Self-RS). FDR-corrected $p < 0.05$ with cluster size $k > 10$ across the masked whole-brain. Cortical atlas regions are labeled based on a combination of parcellations available on GitHub: https://github.com/canlab/Neuroimaging_Pattern_Masks/tree/master/Atlases_and_parcellations/2018_Wager_combined_atlas. L refers the left hemisphere, and R refers to the right hemisphere.

Table S3*Significant clusters of the Other-RS*

Labeled Name	Atlas region (see note)	x	y	z	Voxel number	Max(z)
Superior temporal sulcus (STS)	STSda L	-56	-16	-12	174	0.00053
dIPFC	a9 46v L	-38	50	2	90	0.00085
PCC/Precuneus	Multiple regions	4	-60	34	733	0.00081
IPC / TPJ	PGi L	-52	-60	26	245	0.00048
IPC / TPJ	PGi R	52	-58	30	339	0.00058
Retrosplenial cortex	RSC L	-4	-28	28	15	0.00039
IPC / TPJ	PGs L	-36	-76	42	14	0.00048
IPC / TPJ	PFm L	-42	-58	42	38	0.00034
Anterior insula (AI)	AAIC R	30	12	-10	15	-0.00034
vACC/vmPFC	Multiple regions	-2	40	2	498	-0.00058
Caudate nucleus	Caudate Ca R	20	8	-6	20	-0.00032
Thalamus	Bstem Midbd R	12	-30	-2	24	-0.00049
Anterior insula (AI)	AAIC L	-34	10	-6	13	-0.00027
Ventral striatum	V Striatum L	-2	16	-2	13	-0.00045
Thalamus	Multiple regions	0	-10	10	314	-0.00059
Caudate nucleus	Caudate Ca L	-12	14	6	24	-0.00036
Ventral striatum	Cau R	0	24	8	13	-0.00035
Middle temporal complex (MT+)	LO3 L	-44	-80	12	17	-0.00026
IPC	IP2 L	-46	-40	42	10	-0.00036
Somatosensory cortex	2 L	-30	-44	54	19	-0.00033

Note. Significant positive and negative weights contributing to the Self-Referential Signature (Self-RS). FDR-corrected $p < 0.05$ with cluster size $k > 10$ across the masked whole-brain. Cortical atlas regions are labeled based on a combination of parcellations available on GitHub: https://github.com/canlab/Neuroimaging_Pattern_Masks/tree/master/Atlases_and_parcellations/2018_Wager_combined_atlas. L refers the left hemisphere, and R refers to the right hemisphere.

Table S4*Significant clusters of the MS*

Labeled Name	Atlas region (see note)	x	y	z	Voxel number	Max(z)
Cerebellum	Cblm IX R	2	-54	-44	109	0.00016
Cerebellum	Cblm Crusl R	28	-82	-32	597	0.00040
Cerebellum	Cblm Crusl L	-24	-80	-34	373	0.00028
STS/IFG (incl. vIPFC, AI, STG, temporal pole)	Multiple regions	-48	18	-10	2536	0.00038
Temporal pole	TGd R	50	16	-30	114	0.00014
Orbitofrontal cortex (OFC)	10v R	0	38	-20	15	0.00012
IFG/ventrolateral PFC (vIPFC)	45 R	44	24	-10	195	0.00014
Amygdala	Bstem Ponscd	-24	-12	-10	68	0.00010
Dorsal anterior cingulate cortex (dACC)/vACC (incl. dmPFC, vmPFC, dIPFC, frontal eye fields)	Multiple regions	-6	42	36	3532	0.00036
STS	STSdp R	46	-28	0	20	0.00010
Putamen	Putamen Pa R	30	12	0	13	0.00007
Thalamus	Bstem_SC R	4	-28	2	10	0.00015
Caudate nucleus	Caudate Ca R	16	10	10	107	0.00012
Thalamus	Thal MD	-2	-16	10	96	0.00025
IPC / TPJ	PGi L	-48	-62	26	512	0.00023
Posterior opercular cortex	OP1 L	-50	-24	18	12	0.00008
IPC / TPJ	PGi R	52	-58	30	48	0.00012
PCC	Multiple regions	-2	-50	32	544	0.00023
dIPFC	9p R	18	40	36	42	0.00010
Supplementary motor area (SMA)	8Av L	-36	12	46	68	0.00012
dIPFC	8BL R	14	26	52	19	0.00008
Perirhinal ectorhinal cortex	PeEc L	-34	0	-38	17	-0.00010
Amygdala	Amygdala LB	26	0	-26	13	-0.00012
Cerebellum	Cblm VI L	-36	-42	-24	21	-0.00011
Middle temporal gyrus (MTG) posterior	TE1p R	54	-52	-14	215	-0.00024
MTG anterior	TE2a R	48	-18	-22	27	-0.00012
Hippocampus	H R	20	-16	-20	29	-0.00020
Entorhinal cortex/amygdala	Pir R	36	2	-18	29	-0.00014
Parahippocampal gyrus	PHA3 L	-32	-32	-20	10	-0.00010
Amygdala	Amygdala LB	18	4	-18	29	-0.00015
Posterior temporal cortex/MT+	PH L	-52	-58	-10	419	-0.00029
OFC	11l R	20	54	-14	34	-0.00019
MTG medial	TE1m R	60	-26	-16	12	-0.00011

Posterior temporal cortex	PH R	50	-70	-8	43	-0.00017
Auditory association cortex	A5 R	60	-6	-2	54	-0.00020
STS	STSvp R	56	-40	-4	12	-0.00012
Middle Insula	MI L	-34	14	0	18	-0.00010
Inferior frontal sulcus	IFSa R	48	38	6	57	-0.00016
Extrastriate cortex/MTG	TPOJ2 R	48	-64	12	327	-0.00013
Premotor cortex/Inferior frontal sulcus	6r L	-46	6	26	333	-0.00023
dIPFC	a9 46v L	-36	48	10	17	-0.00016
Extrastriate cortex	V3CD R	36	-84	12	32	-0.00011
IPC	PGp L	-44	-80	16	38	-0.00009
Superior temporal gyrus (STG)	STV R	58	-42	18	35	-0.00012
Supramarginal gyrus/IPC	PF L	-44	-42	44	629	-0.00029
TPJ/Angular gyrus (AG)	PSL R	60	-30	18	18	-0.00011
dIPFC	p9 46v L	-40	30	22	10	-0.00010
Premotor cortex	6v R	50	14	28	32	-0.00015
IPC	IP1 R	30	-64	42	151	-0.00016
IPC/TPJ	PFm R	46	-42	48	576	-0.00033
PCC	5mv L	-10	-38	46	121	-0.00019
Superior parietal cortex/medial intraparietal sulcus	MIP L	-24	-62	50	75	-0.00018
MCC	24dd R	10	-16	42	14	-0.00006
SMA	SCEF R	0	4	52	284	-0.00018
Premotor cortex/SMA	55b R	50	-4	44	31	-0.00013
SMA	6a L	-28	0	50	20	-0.00010
SMA	8Av R	32	30	50	13	-0.00021
SMA	6a L	-24	8	54	23	-0.00012
Primary sensory cortex	1 R	46	-28	62	35	-0.00036

Note. Significant positive and negative weights contributing to the Self-Referential Signature (Self-RS). FDR-corrected $p < 0.05$ with cluster size $k > 10$ across the masked whole-brain. Cortical atlas regions are labeled based on a combination of parcellations available on GitHub: https://github.com/canlab/Neuroimaging_Pattern_Masks/tree/master/Atlases_and_parcellations/2018_Wager_combined_atlas. L refers the left hemisphere, and R refers to the right hemisphere.

Table S5*Prediction results of the mentalizing signatures across training and validation datasets*

Self-Referential Signature (Self-RS)					
Dataset	Study	Sample	n	Classification Task	Prediction Outcome
Training	Study 1	Healthy adult	21	Self vs Other	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.61$
				Self vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.74$
Validation	Study 2	Adolescent	44	Self vs Other	acc.=0.77(+/-0.06), $P=0.0004^{***}$, sens.=0.77, spec.=0.77, AUC=0.88, $d=1.13$
				Self vs Control	acc.=0.98(+/-0.02), $P<0.0001^{***}$, sens.=0.98, spec.=0.98, AUC=1.00, $d=2.60$
Validation	Study 3	Adolescent	61	Self vs Other	acc.=0.75(+/-0.06), $P<0.0001^{***}$, sens.=0.75, spec.=0.75, AUC=0.84, $d=0.95$
				Self vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.52$
Validation	Study 4a	Healthy adult	33	Self vs Other	acc.=0.91(+/-0.05), $P<0.0001^{***}$, sens.=0.91, spec.=0.91, AUC=0.98, $d=1.83$
				Self vs Control	acc.=0.82(+/-0.07), $P=0.0003^{***}$, sens.=0.82, spec.=0.82, AUC=0.93, $d=1.51$
Validation	Study 4b	Schizophrenia patient	23	Self vs Other	acc.=0.78(+/-0.09), $P=0.0106^{***}$, sens.=0.78, spec.=0.78, AUC=0.71, $d=0.50$
				Self vs Control	acc.=0.83(+/-0.08), $P=0.0026^{***}$, sens.=0.83, spec.=0.83, AUC=0.87, $d=1.01$
Validation	Study 5a	Healthy adult	15	Self vs Other	acc.=0.47(+/-0.13), $P>0.20$, sens.=0.47, spec.=0.47, AUC=0.71, $d=0.61$
				Self vs Control	acc.=0.53(+/-0.13), $P>0.20$, sens.=0.53, spec.=0.53, AUC=0.60, $d=0.31$
Validation	Study 5b	Schizophrenia patient	17	Self vs Other	acc.=0.76(+/-0.10), $P=0.0490^{***}$, sens.=0.76, spec.=0.76, AUC=0.78, $d=0.60$
				Self vs Control	acc.=0.71(+/-0.11), $P=0.1435$, sens.=0.71, spec.=0.71, AUC=0.70, $d=0.39$
Validation	Study 5c	Bipolar patient	18	Self vs Other	acc.=0.78(+/-0.10), $P=0.0309^{***}$, sens.=0.78, spec.=0.78, AUC=0.74, $d=0.58$
				Self vs Control	acc.=0.67(+/-0.11), $P>0.20$, sens.=0.67, spec.=0.67, AUC=0.71, $d=0.35$
Other-Referential Signature (Other-RS)					
Signature	Dataset	Sample	n	Classification Task	Prediction Outcome
Training	Study 1	Healthy adult	21	Other vs Self	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.36$
				Other vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.53$
Validation	Study 2	Adolescent	44	Other vs Self	acc.=0.70(+/-0.07), $P=0.0096^{***}$, sens.=0.70, spec.=0.70, AUC=0.83, $d=0.90$
				Other vs Control	acc.=0.95(+/-0.03), $P<0.0001^{***}$, sens.=0.95, spec.=0.95, AUC=0.97, $d=0.99$
Validation	Study 3	Adolescent	61	Other vs Self	acc.=0.75(+/-0.06), $P<0.0001^{***}$, sens.=0.75, spec.=0.75, AUC=0.80, $d=0.81$
				Other vs Control	acc.=0.38(+/-0.06), $P>0.20$, sens.=0.38, spec.=0.38, AUC=0.39, $d=0.39$
Validation	Study 4a	Healthy adult	33	Other vs Self	acc.=0.94(+/-0.04), $P<0.0001^{***}$, sens.=0.94, spec.=0.94, AUC=0.99, $d=2.15$
				Other vs Control	acc.=0.88(+/-0.06), $P<0.0001^{***}$, sens.=0.88, spec.=0.88, AUC=0.95, $d=1.55$

Validation	Study 4b	Schizophrenia patient	23	Other vs Self	acc.=0.70(+/-0.10), $P=0.0931$, sens.=0.70, spec.=0.70, AUC=0.81, $d=0.83$
				Other vs Control	acc.=0.83(+/-0.08), $P=0.0026^{***}$, sens.=0.83, spec.=0.83, AUC=0.90, $d=1.21$
Validation	Study 5a	Healthy adult	15	Other vs Self	acc.=0.87(+/-0.09), $P=0.0074^{***}$, sens.=0.87, spec.=0.87, AUC=0.97, $d=1.65$
				Other vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.73$
Validation	Study 5b	Schizophrenia patient	17	Other vs Self	acc.=0.76(+/-0.10), $P=0.0490^{***}$, sens.=0.76, spec.=0.76, AUC=0.93, $d=1.34$
				Other vs Control	acc.=0.94(+/-0.06), $P=0.0003^{***}$, sens.=0.94, spec.=0.94, AUC=0.99, $d=1.89$
Validation	Study 5c	Bipolar patient	18	Other vs Self	acc.=0.83(+/-0.09), $P=0.0075^{***}$, sens.=0.83, spec.=0.83, AUC=0.86, $d=1.02$
				Other vs Control	acc.=0.94(+/-0.05), $P=0.0002^{***}$, sens.=0.94, spec.=0.94, AUC=0.96, $d=1.79$

Mentalizing Signature (MS)

Signature	Dataset	Sample	n	Classification Task	Prediction Outcome
Training	Study 1	Healthy adult	21	Self vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=5.12$
				Other vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=4.72$
Validation	Study 2	Adolescent	44	Self vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=4.84$
				Other vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=4.99$
Validation	Study 3	Adolescent	61	Self vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.76$
				Other vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.98$
Validation	Study 4a	Healthy adult	33	Self vs Control	acc.=0.97(+/-0.03), $P<0.0001^{***}$, sens.=0.97, spec.=0.97, AUC=1.00, $d=2.13$
				Other vs Control	acc.=0.97(+/-0.03), $P<0.0001^{***}$, sens.=0.97, spec.=0.97, AUC=0.98, $d=1.87$
Validation	Study 4b	Schizophrenia patient	23	Self vs Control	acc.=0.91(+/-0.06), $P<0.0001^{***}$, sens.=0.91, spec.=0.91, AUC=0.99, $d=1.61$
				Other vs Control	acc.=0.96(+/-0.04), $P<0.0001^{***}$, sens.=0.96, spec.=0.96, AUC=0.98, $d=1.74$
Validation	Study 5a	Healthy adult	15	Self vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.47$
				Other vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.36$
Validation	Study 5b	Schizophrenia patient	17	Self vs Control	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=1.95$
				Other vs Control	acc.=0.94(+/-0.06), $P=0.0003^{***}$, sens.=0.94, spec.=0.94, AUC=0.98, $d=1.91$
Validation	Study 5c	Bipolar patient	18	Self vs Control	acc.=0.89(+/-0.07), $P=0.0013^{***}$, sens.=0.89, spec.=0.89, AUC=0.97, $d=1.61$
				Other vs Control	acc.=0.94(+/-0.05), $P=0.0002^{***}$, sens.=0.94, spec.=0.94, AUC=0.99, $d=2.19$

Note. The predictions of signatures were assigned using paired observations with a forced-choice principle. acc. = accuracy; sens. = sensitivity; spec. = specificity, AUC = area under the curve. d refers to the estimated Cohen's d calculated as the mean difference of true and false paired predictions divided by the pooled standard deviation of differences (where difference = input_values[binary_class]- input_values[~binary_class]). Asterisks (***) mark the significant classification accuracies with $p<.05$.

Supplementary Table S6*Prediction results of the ROI classifiers across training and validation datasets*

ROI 1: mPFC			
Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.95(+/-0.05), $P<0.0001^{***}$, sens.=0.95, spec.=0.95, AUC=1.00, $d=2.17$
		Validation (n=211)	acc.=0.66(+/-0.03), $P<0.0001^{***}$, sens.=0.66, spec.=0.66, AUC=0.69, $d=0.37$
	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.29$
		Validation (n=211)	acc.=0.79(+/-0.03), $P<0.0001^{***}$, sens.=0.79, spec.=0.79, AUC=0.88, $d=0.96$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.23$
		Validation (n=211)	acc.=0.63(+/-0.03), $P=0.0002^{***}$, sens.=0.63, spec.=0.63, AUC=0.66, $d=0.36$
	Other vs Control	Training (n=21)	acc.=0.86(+/-0.08), $P=0.0015^{***}$, sens.=0.86, spec.=0.86, AUC=0.97, $d=1.68$
		Validation (n=211)	acc.=0.78(+/-0.03), $P<0.0001^{***}$, sens.=0.78, spec.=0.78, AUC=0.85, $d=0.86$
Predicting mentalizing	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=5.63$
		Validation (n=211)	acc.=0.97(+/-0.01), $P<0.0001^{***}$, sens.=0.97, spec.=0.97, AUC=0.99, $d=1.64$
	Other vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=6.01$
		Validation (n=211)	acc.=0.97(+/-0.01), $P<0.0001^{***}$, sens.=0.97, spec.=0.97, AUC=0.99, $d=1.69$
Differentiating „self“ vs „other“ conditions	Self vs Other	Training (n=21)	acc.=0.95(+/-0.05), $P<0.0001^{***}$, sens.=0.95, spec.=0.95, AUC=1.00, $d=2.10$
		Validation (n=211)	acc.=0.68(+/-0.03), $P<0.0001^{***}$, sens.=0.68, spec.=0.68, AUC=0.68, $d=0.36$
ROI 2: Precuneus/PCC			
Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.81(+/-0.09), $P=0.0072^{***}$, sens.=0.81, spec.=0.81, AUC=0.90, $d=1.20$
		Validation (n=211)	acc.=0.70(+/-0.03), $P<0.0001^{***}$, sens.=0.70, spec.=0.70, AUC=0.74, $d=0.57$
	Self vs Control	Training (n=21)	acc.=0.90(+/-0.06), $P=0.0002^{***}$, sens.=0.90, spec.=0.90, AUC=0.96, $d=1.52$
		Validation (n=211)	acc.=0.68(+/-0.03), $P<0.0001^{***}$, sens.=0.68, spec.=0.68, AUC=0.73, $d=0.59$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.10$
		Validation (n=211)	acc.=0.81(+/-0.03), $P<0.0001^{***}$, sens.=0.81, spec.=0.81, AUC=0.84, $d=0.85$
	Other vs Control	Training (n=21)	acc.=0.95(+/-0.05), $P<0.0001^{***}$, sens.=0.95, spec.=0.95, AUC=1.00, $d=2.72$
		Validation (n=211)	acc.=0.91(+/-0.02), $P<0.0001^{***}$, sens.=0.91, spec.=0.91, AUC=0.96, $d=1.45$
Predicting mentalizing	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.11$
		Validation (n=211)	acc.=0.91(+/-0.02), $P<0.0001^{***}$, sens.=0.91, spec.=0.91, AUC=0.96, $d=1.29$
	Other vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.42$

Differentiating „self“ vs „other“ conditions	Self vs Other	Validation (n=211)	acc.=0.91(+/-0.02), $P<0.0001^{***}$, sens.=0.91, spec.=0.91, AUC=0.97, $d=1.40$
		Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.07$
		Validation (n=211)	acc.=0.75(+/-0.03), $P<0.0001^{***}$, sens.=0.75, spec.=0.75, AUC=0.82, $d=0.78$

ROI 3: TPJ right

Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.81(+/-0.09), $P=0.0072^{***}$, sens.=0.81, spec.=0.81, AUC=0.84, $d=0.69$
		Validation (n=211)	acc.=0.47(+/-0.03), $P>0.20$, sens.=0.47, spec.=0.47, AUC=0.49, $d=0.05$
	Self vs Control	Training (n=21)	acc.=0.76(+/-0.09), $P=0.0266^{***}$, sens.=0.76, spec.=0.76, AUC=0.88, $d=1.03$
		Validation (n=211)	acc.=0.63(+/-0.03), $P=0.0003^{***}$, sens.=0.63, spec.=0.63, AUC=0.69, $d=0.45$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=0.76(+/-0.09), $P=0.0266^{***}$, sens.=0.76, spec.=0.76, AUC=0.86, $d=0.90$
		Validation (n=211)	acc.=0.69(+/-0.03), $P<0.0001^{***}$, sens.=0.69, spec.=0.69, AUC=0.74, $d=0.49$
	Other vs Control	Training (n=21)	acc.=0.81(+/-0.09), $P=0.0072^{***}$, sens.=0.81, spec.=0.81, AUC=0.92, $d=1.12$
		Validation (n=211)	acc.=0.69(+/-0.03), $P<0.0001^{***}$, sens.=0.69, spec.=0.69, AUC=0.74, $d=0.60$
Predicting mentalizing	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.32$
		Validation (n=211)	acc.=0.80(+/-0.03), $P<0.0001^{***}$, sens.=0.80, spec.=0.80, AUC=0.90, $d=1.07$
	Other vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.55$
		Validation (n=211)	acc.=0.86(+/-0.02), $P<0.0001^{***}$, sens.=0.86, spec.=0.86, AUC=0.93, $d=1.15$
Differentiating „self“ vs „other“ conditions	Self vs Other	Training (n=21)	acc.=0.76(+/-0.09), $P=0.0266^{***}$, sens.=0.76, spec.=0.76, AUC=0.84, $d=0.82$
		Validation (n=211)	acc.=0.65(+/-0.03), $P<0.0001^{***}$, sens.=0.65, spec.=0.65, AUC=0.71, $d=0.45$

ROI 4: TPJ left

Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.57(+/-0.11), $P>0.20$, sens.=0.57, spec.=0.57, AUC=0.64, $d=0.23$
		Validation (n=211)	acc.=0.56(+/-0.03), $P=0.0983$, sens.=0.56, spec.=0.56, AUC=0.60, $d=0.24$
	Self vs Control	Training (n=21)	acc.=0.81(+/-0.09), $P=0.0072^{***}$, sens.=0.81, spec.=0.81, AUC=0.75, $d=0.57$
		Validation (n=211)	acc.=0.69(+/-0.03), $P<0.0001^{***}$, sens.=0.69, spec.=0.69, AUC=0.80, $d=0.74$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=0.81(+/-0.09), $P=0.0072^{***}$, sens.=0.81, spec.=0.81, AUC=0.87, $d=0.95$
		Validation (n=211)	acc.=0.71(+/-0.03), $P<0.0001^{***}$, sens.=0.71, spec.=0.71, AUC=0.72, $d=0.44$
	Other vs Control	Training (n=21)	acc.=0.90(+/-0.06), $P=0.0002^{***}$, sens.=0.90, spec.=0.90, AUC=0.98, $d=1.81$
		Validation (n=211)	acc.=0.81(+/-0.03), $P<0.0001^{***}$, sens.=0.81, spec.=0.81, AUC=0.90, $d=1.07$
Predicting mentalizing	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.04$
		Validation (n=211)	acc.=0.89(+/-0.02), $P<0.0001^{***}$, sens.=0.89, spec.=0.89, AUC=0.96, $d=1.19$

Differentiating „self“ vs „other“ conditions	Other vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.70$
		Validation (n=211)	acc.=0.92(+/-0.02), $P<0.0001^{***}$, sens.=0.92, spec.=0.92, AUC=0.98, $d=1.23$
	Self vs Other	Training (n=21)	acc.=0.86(+/-0.08), $P=0.0015^{***}$, sens.=0.86, spec.=0.86, AUC=0.88, $d=1.04$
		Validation (n=211)	acc.=0.68(+/-0.03), $P<0.0001^{***}$, sens.=0.68, spec.=0.68, AUC=0.72, $d=0.45$

ROI 5: aMTG right

Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.67(+/-0.10), $P=0.1893$, sens.=0.67, spec.=0.67, AUC=0.74, $d=0.51$
		Validation (n=211)	acc.=0.64(+/-0.03), $P<0.0001^{***}$, sens.=0.64, spec.=0.64, AUC=0.69, $d=0.47$
	Self vs Control	Training (n=21)	acc.=0.67(+/-0.10), $P=0.1893$, sens.=0.67, spec.=0.67, AUC=0.75, $d=0.59$
		Validation (n=211)	acc.=0.41(+/-0.03), $P>0.20$, sens.=0.41, spec.=0.41, AUC=0.36, $d=-0.37$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=0.81(+/-0.09), $P=0.0072^{***}$, sens.=0.81, spec.=0.81, AUC=0.86, $d=1.11$
		Validation (n=211)	acc.=0.66(+/-0.03), $P<0.0001^{***}$, sens.=0.66, spec.=0.66, AUC=0.69, $d=0.46$
	Other vs Control	Training (n=21)	acc.=0.81(+/-0.09), $P=0.0072^{***}$, sens.=0.81, spec.=0.81, AUC=0.94, $d=1.35$
		Validation (n=211)	acc.=0.84(+/-0.03), $P<0.0001^{***}$, sens.=0.84, spec.=0.84, AUC=0.92, $d=1.13$
Predicting mentalizing	Self vs Control	Training (n=21)	acc.=0.76(+/-0.09), $P=0.0266^{***}$, sens.=0.76, spec.=0.76, AUC=0.92, $d=1.25$
		Validation (n=211)	acc.=0.80(+/-0.03), $P<0.0001^{***}$, sens.=0.80, spec.=0.80, AUC=0.87, $d=1.00$
	Other vs Control	Training (n=21)	acc.=0.86(+/-0.08), $P=0.0015^{***}$, sens.=0.86, spec.=0.86, AUC=0.95, $d=1.39$
		Validation (n=211)	acc.=0.83(+/-0.03), $P<0.0001^{***}$, sens.=0.83, spec.=0.83, AUC=0.89, $d=1.00$
Differentiating „self“ vs „other“ conditions	Self vs Other	Training (n=21)	acc.=0.81(+/-0.09), $P=0.0072^{***}$, sens.=0.81, spec.=0.81, AUC=0.84, $d=0.90$
		Validation (n=211)	acc.=0.65(+/-0.03), $P<0.0001^{***}$, sens.=0.65, spec.=0.65, AUC=0.71, $d=0.51$

ROI 6: aMTG left

Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.95(+/-0.05), $P<0.0001^{***}$, sens.=0.95, spec.=0.95, AUC=0.98, $d=1.83$
		Validation (n=211)	acc.=0.57(+/-0.03), $P=0.0386^{***}$, sens.=0.57, spec.=0.57, AUC=0.63, $d=0.35$
	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.53$
		Validation (n=211)	acc.=0.75(+/-0.03), $P<0.0001^{***}$, sens.=0.75, spec.=0.75, AUC=0.81, $d=0.87$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=0.86(+/-0.08), $P=0.0015^{***}$, sens.=0.86, spec.=0.86, AUC=0.96, $d=1.69$
		Validation (n=211)	acc.=0.71(+/-0.03), $P<0.0001^{***}$, sens.=0.71, spec.=0.71, AUC=0.80, $d=0.81$
	Other vs Control	Training (n=21)	acc.=0.95(+/-0.05), $P<0.0001^{***}$, sens.=0.95, spec.=0.95, AUC=0.97, $d=1.71$
		Validation (n=211)	acc.=0.73(+/-0.03), $P<0.0001^{***}$, sens.=0.73, spec.=0.73, AUC=0.80, $d=0.74$
Predicting mentalizing	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=4.43$

Differentiating „self“ vs „other“ conditions	Other vs Control	Validation (n=211)	acc.=0.92(+/-0.02), $P<0.0001^{***}$, sens.=0.92, spec.=0.92, AUC=0.97, $d=1.30$
		Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.49$
	Self vs Other	Validation (n=211)	acc.=0.94(+/-0.02), $P<0.0001^{***}$, sens.=0.94, spec.=0.94, AUC=0.98, $d=1.47$
		Training (n=21)	acc.=0.95(+/-0.05), $P<0.0001^{***}$, sens.=0.95, spec.=0.95, AUC=0.99, $d=1.91$
		Validation (n=211)	acc.=0.64(+/-0.03), $P<0.0001^{***}$, sens.=0.64, spec.=0.64, AUC=0.69, $d=0.49$

ROI 7: SMA left

Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.67(+/-0.10), $P=0.1893$, sens.=0.67, spec.=0.67, AUC=0.70, $d=0.23$
		Validation (n=211)	acc.=0.52(+/-0.03), $P>0.20$, sens.=0.52, spec.=0.52, AUC=0.51, $d=-0.02$
	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.44$
		Validation (n=211)	acc.=0.48(+/-0.03), $P>0.20$, sens.=0.48, spec.=0.48, AUC=0.53, $d=0.12$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=0.43(+/-0.11), $P>0.20$, sens.=0.43, spec.=0.43, AUC=0.51, $d=-0.02$
		Validation (n=211)	acc.=0.50(+/-0.03), $P>0.20$, sens.=0.50, spec.=0.50, AUC=0.51, $d=-0.08$
	Other vs Control	Training (n=21)	acc.=0.62(+/-0.11), $P>0.20$, sens.=0.62, spec.=0.62, AUC=0.81, $d=0.87$
		Validation (n=211)	acc.=0.71(+/-0.03), $P<0.0001^{***}$, sens.=0.71, spec.=0.71, AUC=0.75, $d=0.60$
Predicting mentalizing	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=4.30$
		Validation (n=211)	acc.=0.85(+/-0.02), $P<0.0001^{***}$, sens.=0.85, spec.=0.85, AUC=0.94, $d=1.20$
	Other vs Control	Training (n=21)	acc.=0.95(+/-0.05), $P<0.0001^{***}$, sens.=0.95, spec.=0.95, AUC=1.00, $d=3.05$
		Validation (n=211)	acc.=0.86(+/-0.02), $P<0.0001^{***}$, sens.=0.86, spec.=0.86, AUC=0.93, $d=1.20$
Differentiating „self“ vs „other“ conditions	Self vs Other	Training (n=21)	acc.=0.62(+/-0.11), $P>0.20$, sens.=0.62, spec.=0.62, AUC=0.69, $d=0.20$
		Validation (n=211)	acc.=0.50(+/-0.03), $P>0.20$, sens.=0.50, spec.=0.50, AUC=0.50, $d=-0.03$

ROI 8: Cerebellum bilateral

Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.62(+/-0.11), $P>0.20$, sens.=0.62, spec.=0.62, AUC=0.65, $d=0.29$
		Validation (n=211)	acc.=0.52(+/-0.03), $P>0.20$, sens.=0.52, spec.=0.52, AUC=0.56, $d=0.10$
	Self vs Control	Training (n=21)	acc.=0.86(+/-0.08), $P=0.0015^{***}$, sens.=0.86, spec.=0.86, AUC=0.94, $d=1.48$
		Validation (n=211)	acc.=0.67(+/-0.03), $P<0.0001^{***}$, sens.=0.67, spec.=0.67, AUC=0.71, $d=0.39$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=0.48(+/-0.11), $P>0.20$, sens.=0.48, spec.=0.48, AUC=0.54, $d=0.06$
		Validation (n=211)	acc.=0.54(+/-0.03), $P>0.20$, sens.=0.54, spec.=0.54, AUC=0.55, $d=0.05$
	Other vs Control	Training (n=21)	acc.=0.76(+/-0.09), $P=0.0266^{***}$, sens.=0.76, spec.=0.76, AUC=0.85, $d=0.81$
		Validation (n=211)	acc.=0.45(+/-0.03), $P>0.20$, sens.=0.45, spec.=0.45, AUC=0.46, $d=-0.10$

Predicting mentalizing	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=2.70$
		Validation (n=211)	acc.=0.69(+/-0.03), $P<0.0001^{***}$, sens.=0.69, spec.=0.69, AUC=0.76, $d=0.61$
	Other vs Control	Training (n=21)	acc.=0.95(+/-0.05), $P<0.0001^{***}$, sens.=0.95, spec.=0.95, AUC=1.00, $d=2.61$
		Validation (n=211)	acc.=0.73(+/-0.03), $P<0.0001^{***}$, sens.=0.73, spec.=0.73, AUC=0.77, $d=0.68$
Differentiating „self“ vs „other“ conditions	Self vs Other	Training (n=21)	acc.=0.62(+/-0.11), $P>0.20$, sens.=0.62, spec.=0.62, AUC=0.64, $d=0.26$
		Validation (n=211)	acc.=0.54(+/-0.03), $P>0.20$, sens.=0.54, spec.=0.54, AUC=0.56, $d=0.07$

ROI 9: Cerebellum right

Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.48(+/-0.11), $P>0.20$, sens.=0.48, spec.=0.48, AUC=0.46, $d=-0.31$
		Validation (n=211)	acc.=0.40(+/-0.03), $P>0.20$, sens.=0.40, spec.=0.40, AUC=0.37, $d=-0.34$
	Self vs Control	Training (n=21)	acc.=0.76(+/-0.09), $P=0.0266^{***}$, sens.=0.76, spec.=0.76, AUC=0.84, $d=0.95$
		Validation (n=211)	acc.=0.73(+/-0.03), $P<0.0001^{***}$, sens.=0.73, spec.=0.73, AUC=0.81, $d=0.79$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=0.67(+/-0.10), $P=0.1893$, sens.=0.67, spec.=0.67, AUC=0.62, $d=0.21$
		Validation (n=211)	acc.=0.49(+/-0.03), $P>0.20$, sens.=0.49, spec.=0.49, AUC=0.47, $d=-0.10$
	Other vs Control	Training (n=21)	acc.=0.90(+/-0.06), $P=0.0002^{***}$, sens.=0.90, spec.=0.90, AUC=0.95, $d=1.31$
		Validation (n=211)	acc.=0.60(+/-0.03), $P=0.0058^{***}$, sens.=0.60, spec.=0.60, AUC=0.64, $d=0.38$
Predicting mentalizing	Self vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.52$
		Validation (n=211)	acc.=0.88(+/-0.02), $P<0.0001^{***}$, sens.=0.88, spec.=0.88, AUC=0.96, $d=1.32$
	Other vs Control	Training (n=21)	acc.=1.00(+/-0.00), $P<0.0001^{***}$, sens.=1.00, spec.=1.00, AUC=1.00, $d=3.61$
		Validation (n=211)	acc.=0.88(+/-0.02), $P<0.0001^{***}$, sens.=0.88, spec.=0.88, AUC=0.95, $d=1.36$
Differentiating „self“ vs „other“ conditions	Self vs Other	Training (n=21)	acc.=0.57(+/-0.11), $P>0.20$, sens.=0.57, spec.=0.57, AUC=0.51, $d=-0.14$
		Validation (n=211)	acc.=0.43(+/-0.03), $P>0.20$, sens.=0.43, spec.=0.43, AUC=0.41, $d=-0.24$

ROI 10: Cerebellum left

Trained for	Task	Sample	Prediction Outcome
Predicting Self-mentalizing	Self vs Other	Training (n=21)	acc.=0.29(+/-0.10), $P>0.20$, sens.=0.29, spec.=0.29, AUC=0.45, $d=-0.05$
		Validation (n=211)	acc.=0.49(+/-0.03), $P>0.20$, sens.=0.49, spec.=0.49, AUC=0.50, $d=-0.02$
	Self vs Control	Training (n=21)	acc.=0.76(+/-0.09), $P=0.0266^{***}$, sens.=0.76, spec.=0.76, AUC=0.90, $d=0.94$
		Validation (n=211)	acc.=0.71(+/-0.03), $P<0.0001^{***}$, sens.=0.71, spec.=0.71, AUC=0.78, $d=0.73$
Predicting Other-mentalizing	Other vs Self	Training (n=21)	acc.=0.62(+/-0.11), $P>0.20$, sens.=0.62, spec.=0.62, AUC=0.63, $d=0.12$
		Validation (n=211)	acc.=0.52(+/-0.03), $P>0.20$, sens.=0.52, spec.=0.52, AUC=0.50, $d=-0.02$
	Other vs Control	Training (n=21)	acc.=0.81(+/-0.09), $P=0.0072^{***}$, sens.=0.81, spec.=0.81, AUC=0.88, $d=1.09$

Predicting mentalizing	Self vs Control	Validation (n=211)	acc.=0.62(+/-0.03), $P=0.0006^{***}$, sens.=0.62, spec.=0.62, AUC=0.64, $d=0.35$
		Training (n=21)	acc.=0.90(+/-0.06), $P=0.0002^{***}$, sens.=0.90, spec.=0.90, AUC=0.99, $d=2.26$
	Other vs Control	Validation (n=211)	acc.=0.85(+/-0.02), $P<0.0001^{***}$, sens.=0.85, spec.=0.85, AUC=0.93, $d=1.12$
		Training (n=21)	acc.=0.90(+/-0.06), $P=0.0002^{***}$, sens.=0.90, spec.=0.90, AUC=0.99, $d=2.10$
Differentiating „self“ vs „other“ conditions	Self vs Other	Validation (n=211)	acc.=0.86(+/-0.02), $P<0.0001^{***}$, sens.=0.86, spec.=0.86, AUC=0.93, $d=1.08$
		Training (n=21)	acc.=0.48(+/-0.11), $P>0.20$, sens.=0.48, spec.=0.48, AUC=0.58, $d=0.11$
		Validation (n=211)	acc.=0.48(+/-0.03), $P>0.20$, sens.=0.48, spec.=0.48, AUC=0.49, $d=-0.04$

Note. Each ROI was trained for four classification tasks and tested in validation datasets. acc. = accuracy; sens. = sensitivity, spec. = specificity, AUC = area under the curve. d refers to the estimated *Cohen's d* calculated as the mean difference of true and false paired predictions divided by the pooled standard deviation of differences (where difference = input_values[binary_class]- input_values[~binary_class]). Asterisks (***) mark the significant classification accuracies with $p<.05$.
