ELISA MELLONI
IRCCS San Raffaele Scientific Institute, Vita-Salute San Raffaele University
melloni.elisa@hsr.it

FRANCESCO BENEDETTI
IRCCS San Raffaele Scientific Institute, Vita-Salute San Raffaele University
benedetti.francesco@hsr.it

BENEDETTA VAI
IRCCS San Raffaele Scientific Institute, Vita-Salute San Raffaele University
Fondazione Centro San Raffaele
vai.benedetta@hsr.it

ELISABETTA LALUMERA
Milano-Bicocca University
elisabetta.lalumera@unimib.it

NOT UNDERSTANDING OTHERS. THE RDOC APPROACH TO THEORY OF MIND AND EMPATHY DEFICITS IN SCHIZOPHRENIA, BORDERLINE PERSONALITY DISORDER AND MOOD DISORDERS

abstract

The Research Domani Criteria framework (RdoC) encourages research on specific impairments present across traditional nosological categories and suggests a list of biological and behavioral measures for assessing them. After a description of RdoC, in this article we focus on impairments of the ability of understanding others, specifically in Theory of Mind and empathy. We illustrate recent evidence on brain anomalies correlating with these deficits in Schizophrenia, Addiction Disorders and Mood Disorders populations. In the last section, we zoom out and consider this kind of research vis-à-vis the objection of being reductionistic that is, in favoring mechanistic accounts of mental disorders. We argue that metaphysical reductionism and explanatory reductionism are not conceptually entailed by the RdoC framework.

keywords

schizophrenia, addiction disorders, mood disorders, theory of mind, empathy, rdoc, reductionism, neuroimaging

Phenomenology and Mind, n. 18 - 2020, pp. 162-181
DOI: 10.17454/pam-1813
Web: http://www.rosenbergesellier.it/elenco-libri?aaidriv=14

ISSN 2280-7853 (print) - ISSN 2239-4028 (on line)

1. Introduction

Our social life, the possibility of having successful and fulfilling relations and exchanges with other people, depends on our capacity to understand their actions and emotions, and to adjust our behaviour accordingly. Alterations of such capacity are associated with a wide range of disabling and distressing mental conditions, including autism but also schizophrenia, mood disorders and substance abuse disorders.>
(Cotter *et al.*, 2018)

Psychology has coined two concepts for the capacity of Understanding Others, namely Theory of mind and Empathy, both grounded in philosophical tradition, and now employed in neuropsychology. In this article we illustrate their use as constructs in contemporary psychiatric research. One way to approach the problem of not understanding others is to study the variations of neural and brain correlates of Empathy and Theory of Mind, in experimental paradigms using neuroimaging techniques with behavioural or self - report controlled variables. The guiding hypothesis is that dysfunctions in the social cognitive mechanisms of the brain can become clinical markers for more precise diagnosis and treatment of people's impairments in social life. The first goal of this article is to provide a short narrative review of current studies, focusing on schizophrenia, drug abuse and mood disorders.
The general framework for this kind of studies is the RdoC (Research Domain Criteria) Project, launched in 2009 by the National Institute of Mental Health (NIMH) of the United States in order to "develop, for research purposes, new ways of classifying mental disorders based on dimensions of observable behavior and neuroblological measures" (Morris and Cuthbert, 2012). RdoC favours a bottom-up approach to research in psychopathology. Rather than considering traditional nosological categories, such as Schizophrenia or major Depression, individuated by signs and symptoms, and searching for the underlying pathophysiological mechanisms, it starts with selecting out broad domains of functioning of the human mind (eg., Cognitive processes and Social Processes) each containing subordinate constructs (e.g., Attention, Social Communication), which can be analyzed at different levels (e.g. brain circuits, behavior, self-reports). In the RdoC project, Theory of Mind and Empathy are precisely two subconstructs of Social Processes.
In the ten years from its launch, RdoC actively functioned as a matrix for new research, as we will see in the next sections. However, it also sparked a debate on both its structural characteristics (i.e. choice of domains and constructs) (eg. Lilienfeld and Treadway, 2016; Hofmann and Zachar, 2017), and on conceptual assumptions about the nature of mental

disorders[1]. In particular, it has been described as having a reductionist approach to mental disorders, as it identifies them with brain dysfunctions - a psychiatry without psyche (see eg. Parnas, 2014). The second goal of this article is to address, albeit briefly, the reductionist objection. As others have noted, RdoC has been often misunderstood on this issue especially in early comments (Cuthbert and Kozak, 2013; Faucher and Goyer, 2015; Lake *et al.*, 2017). In particular, we will argue that the objection can be blocked by pointing out that brain-behaviour correlations need not be taken as metaphysical identities, and moreover, even granting that metaphysical reductionism were assumed, that would still be compatible with the view that the explanation of any mental disorder or symptom ought to be multi-level. In fact, the methodological problem of RdoC, if any, is not reductionism, but dealing with a heterogeneity of constructs that call for the integration of different measures.
This is how the article is organized. Section 2 is dedicated to an introduction to the RdoC framework. Section 3 briefly illustrates the role of ToM and empathy as psychological constructs employed in brain studies. Sections 4, 5, and 6 show how similar ToM and empathy deficits reveal in three psychiatric conditions which are extremely heterogeneous in their clinical pictures: Schizophrenia and Mood Disorders, which reflect the Kraepelinian dichotomy between the major endogenous psychoses (Kraepelin, 1913), and which are still mutually excluding diagnoses in modern psychiatry, as defined by not-overlapping criteria in the DSM-5 classification (A.P.A., 2013); and Borderline Personality Disorder. Section 7 deals with the philosophical reductionist objection, and Section 8 is for the concluding remarks.

2. The RdoC framework

RdoC provides a matrix for organizing research, publicly accessible online (NIMH 2019). At the rows of the matrix are 6 domains of functioning: Negative Valence, Positive Valence, Cognitive Systems, Systems for Social Processes, Arousal/Modulatory Systems, and Sensorimotor Systems. Each domain is specified by subordinate ones, called "constructs". At the column of the RdoC matrix are the "units of analysis", the six different classes of variables that can be used to measure constructs and subconstructs. They are Genes, Molecules, Neural Circuits, Physiology, Behavior, and Self-Report (including patient verbal report). Circuits can refer to measurements of particular circuits as studied by neuroimaging techniques, and/or other measures validated by animal models or functional neuroimaging (e.g., emotion-modulated startle, event-related potentials with established source localization). Physiology refers to measures that are well-established indices of certain constructs, but that do necessarily not tap circuits directly (e.g., heart rate, cortisol). Behavior can refer variously to behavioral tasks or to systematic behavioral observations (e.g., a working memory task, a toddler behavioral assessment), while self-reports are interview-based scales, self-report questionnaires, or other instruments that may encompass normal-range and/or abnormal aspects of the dimension of interest. The word "construct" is here used in the typical psychological meaning, indicating a hypothetical entity, often not observable, which serves to organize a set of data (Cronbach and Meele, 1955). As for "units of analysis", the expression was preferred to "levels" as the latter suggested reduction on one kind of variable to the other (Stanislow *et al.* 2020). We will return to the issue of RdoC and reductionism in Section 7 below.
RdoC can be seen as a response to major criticisms directed at the two most widely used nosologies, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5, APA 2013) and the International Classification of Diseases (ICD-11, WHO 2018), at least when employed as

1 We acknowledge that the conditions that some consider psychopathology or disorder, others consider neuro- or mental diversity. Since our article concern RDoC specifically, we will adopt the preferred terminology of the NIMH and APA, favouring "disorder" and "psychopathology", with no intention of arguing against the neurodiversity movement.

research blueprints. The first criticism concerns the possibility of integration with psychiatric research. Neuroscience, genetics, biology of the brain, and neuroimaging have undergone a tremendous progress in the last 30 years, but so far, they have failed to have a direct and significant impact on the diagnosis and treatment behavioural difficulties (eg. Casey *et al.* 2013). Critics within the discipline have long been pointing out that the two most used nosologies, namely the International Classification of Diseases, developed by the World Health Organization, and the Diagnostic and Statistical Manual of Mental Disorders, intrinsically do not favour the integration of psychiatric science and clinic, when used in research design. As DSM-5 and ICD-10 identify disorders for diagnostic purposes by signs and symptoms alone, the same diagnosis can happen to be applied to people with very different psychological and brain conditions, thus failing to reflect genuine uniformities of mechanisms at the psychological, neuroscientific, biological and genetic level (Hyman, 2007, 2010, Insel *et al.* 2010; Nesse and Stein, 2012; Stanislow *et al.*, 2010). The same issue has been raised by philosophers of psychiatry, who actively participate in the foundational debate on the psychiatric paradigm (Murphy, 2006; Zachar *et al.*, 2014; Tabb, 2015; Tsou, 2015). The research units individuated by the RdoC matrics, eg. subconstructs, are more fine-grained and homogeneous than traditional nosological categories, in order to facilitate integration of basic science findings with clinical psychiatry.

The second issue raised by critics of the DSM-5 concerns its categorical nature. DSM allows yes/no diagnoses only, and people who do not meet the required diagnostic criteria either do not receive a diagnosis or fall under the "not otherwise specified" categories. Categoricity is partially connected with the aim of the manual, namely, to classify patients for the needs of healthcare funding agencies, such as MediCare in the US, and thereby to facilitate decisions about who should be treated or reimbursed for their mental condition. Critics, however, have long been claiming that categoricity is unfit for research purposes, as it forces clear-cut boundaries on what appears to be a continuum of conditions from typical to atypical, or functional to dysfunctional behavior (eg. Cuthbert, 2014; London, 2014; Yee, 2015). Differently, in the RdoC framework - as we read on the official website - "constructs are studied along a span of functioning from normal to abnormal with the understanding that each is situated in, and affected by, environmental and neurodevelopmental contexts" (NIMH 2019).

The formation process of RdoC involved consensus meetings of experts (working groups), and its stages were made relatively open to the public - a procedure quite similar to that leading to the publication DSM-5, and common in scientific psychiatry (eg. Barch *et al.*, 2009; Stanislow *et al.* 2010). Each working group was assigned a domain and had to reach consensus on the constructs and subconstructs of such domain, and on their definitions. The criteria employed for individuating and defining the constructs were the following: sufficient evidence for its validity as a functional unity of behavior or cognitive process, sufficient evidence for a neural circuit or symptom that played a primary role in implementing the constructs function; and sufficient relevance for understanding some aspects of psychopathology (Cuthbert, 2015). However, the matrix is meant to be a work in progress, and constructs and even domains may change as new evidence is being considered. For example, the Sensorimotor System Domain was added in 2018 (Harrison *et al.*, 2019). It is important that RdoC research unities do not suffer the same fate as DSM and ICD ones, in becoming reified and resistant to change (Charney, *et al.*,2013).

Let us focus now on the Social Processes Domain more closely, in order to introduce the case studies of this article. The NIMH RDoC Working Group for the Social Processes Domain reached consensus on the definitions of four constructs: Affiliation and Attachment, Social Communication Perception and Understanding of Self, and Perception and Understanding of Others. The latter is defined as "The processes and/or representations involved in being

aware of, accessing knowledge about, reasoning about, and/or making judgments about other animate entities, including information about cognitive or emotional states, traits or abilities". It was organized into the following sub-constructs: Animacy Perception, the ability to appropriately perceive that another entity is an agent (i.e., has a face, interacts contingently, and exhibits biological motion); Action Perception, the ability to perceive the purpose of an action being performed by an animate entity; and Understanding Mental States, the ability to make judgments and/or attributions about the mental state of other animate entities that allows one to predict or interpret their behaviors. Here "mental state" refers to intentions, beliefs, desires, and emotions. As said above, the concepts needed to operationalize such ability come from developmental psychology and fall under the umbrella term "Theory of Mind" (ToM) and empathy. In the next section we briefly recap the main features of ToM and empathy from a neuroscience perspective, before proceeding to psychopathology research in section 4.

3. Theory of Mind and Empathy

Human social life critically depends on understanding the internal causes of behavior. ToM ability is considered to be crucial for successful social interactions (Samson *et al.*, 2007) and it is involved in affect regulation, impulse control and self-monitoring (Fonagy and Target, 1996). Two different subcomponent of ToM have been identified: the cognitive ToM (cToM) primarily concerns the representations of others' knowledge, beliefs, intentions and other neutral states, while the affective ToM (aToM) is involved in others' affective states (i.e., emotions, feelings and desires) and is supposed to require the appreciation of the emotional states of others. (Brothers and Ring, 1992; Shamay-Tsoory, 2011).

ToM deficits are key predictor of social function outcomes, mental health and quality of life (Milders *et al.*, 2006) and have been suggested to exert a crucial role in the developing and maintenance of different psychopathologies such as autism, anorexia, schizophrenia (SZ), borderline personality disorder (BPD) and bipolar disorder (BD) (Bora, 2009, Bora *et al.*, 2016, Lazarus *et al.*, 2014, Leppanen *et al.*, 2018).

A large body of research indicates that ToM starts in the first few years, in typically developing children, but continues to evolve across childhood and adolescence: environmental experiences and biological maturation of involved brain regions co-occur and interact in determining ToM development (Wolf *et al.*, 2010). In addition, children's developing language abilities (Milligan *et al.*, 2007) and executive functions (Moses *et al.*, 2010) play an important role in the acquisition of socio-cognitive skills. At the age of about 18–24 months, typically-developing children manifest self-awareness passing successfully the mirror test, where the ability of visual self-recognition is evaluated (Perner and Davies, 1991; Povinelli, 1993). The child is now able to distinguish between the representation of a real event and the representation of an hypothetical state, such as a thought, and thus starts to pretend and simulate in fictional game (Leslie and Keeble, 1987). Until the age of 3-4 years, a child is not able to explicitly distinguish between his or her own and others' beliefs, and so understand that someone may hold false beliefs, *i.e.* representations about the world that may contrast with reality (Perner *et al.*, 2011). The next step in ToM acquisition occurs at an age of about six or seven years, when children are able to use second-order attributions, involving the understanding of one person's beliefs about another person's beliefs about reality (Perner and Winner, 1985). At this age also metaphor, irony and the ability to reliably distinguish jokes from lies begins to mature (Gross and Harris, 1988, Sullivan *et al.*, 1995) This rather late development might indicate that children at first have to gain an understanding of others' beliefs in order to appreciate that beliefs guide others' emotions (Adrian *et al.*, 2005). Even more complex is the comprehension of a *faux pas* situation, which happens when someone unintentionally says or does something inappropriate in a social situation. Children

presumably start to understand faux pas from the age 9 to 11 years (Baron-Cohen *et al.*, 1999). This kind of situation has been demonstrated to rely on to both an affective ToM and second-order ToM (Baron-Cohen *et al.*, 1999, Meristo *et al.*, 2007), suggesting that affective ToM might have a prolonged developmental trajectory compared to cognitive subcomponent.
Several brain areas have been implicated in ToM processes. Functional neuroimaging studies on healthy subjects have found that during ToM reasoning a neural network that encompasses superior temporal sulcus (STS), temporal parietal junction (TPJ), Precuneus (PCun), posterior cingulate cortex (PCC) and medial prefrontal cortex (MPFC) (Van Overwalle and Baetens, 2009; Vogeley *et al.*, 2001). Less frequently, the amygdala (Amy) and the anterior temporal lobe have also been associated to ToM (Frith and Frith, 2007; Mar, 2011). A "core" ToM network, activated whenever we are reasoning about mental states, across a large range of tasks and stimulus-formats, seems to involve at least the MPFC, and the TPJ (Schurz *et al.*, 2014).
A key dimension which emerges from social cognition literature is the distinction between low-level perceptual processes necessary to decode social information from the environment, and higher-level cognitive processes which integrate and interpret it. Several studies have identified differential neural underpinnings of these different levels. The implicit automated operations (e.g. decoding facial expressions and biological motion) recruit fusiform face area, STS, inferior frontal gyrus (IFG), and premotor areas (Dapretto *et al.*, 2006; Malhi *et al.*, 2008), whereas the more demanding explicit mental state reasoning recruits the MPFC and the TPJ (Amodio and Frith, 2006; Saxe and Wexler, 2005).
There is emerging evidence that social cognitive deficits may represent a transdiagnostic issue, potentially serving as a marker of neurological abnormality across a wide range of clinical conditions (Bora *et al.*, 2016; Bora and Pantelis, 2016; Domes *et al.*, 2009; Kohler *et al.*, 2011; Kohler *et al.*, 2010). In a systematic review of meta analyses significant deficits in the ability to identify emotions from facial expressions and to successfully complete ToM tasks have been observed in several different developmental, neurological and psychiatric disorders, starting from early stages and getting more severe on conditions with longer disease duration (Cotter *et al.*, 2018). Described below are some example of how social cognition disabilities can shed light on different clinical conditions and potentially guide new-targeted treatments.

4. Schizophrenia

Social cognition abilities are severely impaired in schizophrenia (SZ) (Corcoran *et al.*, 1997, Kington *et al.*, 2000) and ToM in particular have been considered a stronger predictor of functional outcome than other social cognitive measures (Fett *et al.*, 2011). Evidence suggest specific deficit in aToM (Herold *et al.*, 2002; Kern *et al.*, 2009; Mo *et al.*, 2008; Shamay-Tsoory *et al.*, 2007), but impairments in cToM have also been repeatedly identified (Greig *et al.*, 2004; Horan *et al.*, 2009; Inoue *et al.*, 2006; Kelemen *et al.*, 2005). Several studies showed that SZ is characterized by abnormal neural response in areas deeply involved in ToM, such as MPFC, TPJ, STS (Brunet *et al.*, 2003; Ciaramidaro *et al.*, 2015; Walter *et al.*, 2009). However, results are still heterogeneous in terms of increased or decreased neural activation compared to healthy controls, especially for PFC and temporal regions. This is probably due to an intrinsic complexity related both to the disease, and to ToM itself, which requires the integration of several abilities (Bosia *et al.*, 2012). For example, the so-called positive symptoms of SZ (*i.e.* delusions, hallucinations and unusual or disorganized behavior) could be related to increased activation in the sensory and perceptual aspects of ToM, such as STS and SCX, but they could also be attributed to a reduced neural response in MPFC, the area involved in integrating this information with higher cognitive functions (Martin *et al.*, 2014). Paranoid symptoms of SZ have been linked to hyperactivity in the MPFC and TPJ/STS in ToM control conditions or in non-social situations; this could be related to an hyperactive intention detector that fails to deactivate when viewing socially neutral or intention-free scenes (Backasch *et al.*,

2013; Ciaramidaro *et al.*, 2015; Martin *et al.*, 2014; Walter *et al.*, 2009). On the other hand, hypoactivation of insula, thalamus and striatum during ToM task have been associated to reduced self-reference and awareness that could contribute to passivity symptoms, such as third-person auditory hallucinations or delusion of control (Brune *et al.*, 2008). A recent meta-analysis showed consistent neural correlates of ToM impairment in SZ: under-activation was identified in the MPFC, left orbito-frontal cortex, and in a small section of the left posterior temporo-parietal junction, while robust over-activation was identified in a more dorsal, bilateral section of the TPJ (Kronbichler *et al.*, 2017). In conclusion, SZ patients show less specialized brain activation in regions linked to ToM and increased activation in attention-related networks suggesting compensatory effects. Evidence showed that impairments in ToM ability develops across adolescence in young people with high clinical risk for psychosis, while it presents a normal age-related trajectory under antipsychotic and antidepressant medication (Davidson *et al.*, 2018). These processes seem to parallel a progressive gray matter reduction in superior temporal gyrus that precedes the first episode of psychosis and correlates with the severity of delusions at follow-up (Takahashi *et al.*, 2009). Additional extensive losses of both gray and white matter in lateral fronto-temporal regions and left anterior cingulate gyrus were observed over time (Farrow *et al.*, 2005) and functional and structural abnormalities in the same areas were associated with deficits in performance on tasks targeting ToM and empathy in SZ patients (Benedetti *et al.*, 2009).
A similar pattern of grey matter pathology, reduced performance and greater recruiting of neuronal resources has been observed in prefrontal cortex of SZ patients with working memory tasks, and parallel improvements in performance and reductions in neural activations have been reported after successful treatment (Callicott *et al.*, 2000). A similar "physiological inefficiency" could link the ToM and EMP deficits in schizophrenia with abnormal structure and function of the posterior temporal lobe, with the schizophrenic process specifically targeting these areas soon at the beginning of the illness and during its early course (Benedetti *et al.*, 2009). Thus, it can be surmised that a condition involving a general pattern of brain structural and functional abnormalities will impact human behavior with many subtle changes, including ToM and empathy deficits as defined according to the RDoC perspective; and that these single characteristics can be revealed one by one in experimental settings specifically designed to target them, and which all together lead to the multi-faceted phenotype of clinical Schizophrenia.

5. Borderline personality disorder

Borderline personality disorder (BPD) is a severe psychiatric condition, characterized by a marked instability in affect regulation, impulse control, social cognition skills and interpersonal relationships. Research reported an overall deficit in emotion recognition accuracy, especially for the negative emotions of disgust and anger (Unoka *et al.*, 2011). However, the largest deficit was observed for the identification of neutral facial expressions, suggesting that BPD patients tend to misattribute emotions, mainly negative, to faces that do not convey emotional information (Mitchell *et al.*, 2014). Previous studies which assessed ToM in BPD have yielded contrasting results, including impaired (Ghiassi *et al.*, 2010; Scott *et al.*, 2011), preserved (Murphy, 2006; Schilling *et al.*, 2012), or even superior abilities (Arntz *et al.*, 2009; Fertuck *et al.*, 2009; Franzen *et al.*, 2011) compared to healthy controls. This performance variability seems to be largely dependent on task demand in terms of ToM processing target (emotion or mental-state recognition and intentional attribution). A recent meta-analysis on more than 400 studies, pointed out the BPD patients are significantly impaired in their overall ToM capacities compared to HC, but with a relatively small effect-size. In particular, a poorer performance in mental state reasoning, such as faux pas, was found, in contrast to a relatively intact affective decoding and discriminating capacities (Nemeth *et al.*, 2018). A largely used

task to test brain regions involved in mental states decoding is the 'Reading the Mind in the Eyes task' (RMET) developed by Baron-Cohen and colleagues (2001). In this task, participants are required to match the mental state of a person, shown in a photograph of their eye regions, with one of four mental state words. Research on healthy controls frequently report a significant activation in inferior frontal gyrus (IFG) and middle temporal gyrus extending to posterior superior temporal sulcus (pSTS) during RMET task (Thye *et al.*, 2018). Compared to controls, BPD patients seem to differently activate the dedicated social brain networks. Indeed, decreased STS/STG and enhanced insula responses have been associated to reduced ToM abilities in BPD patients (Frick *et al.*, 2012; Mier *et al.*, 2013) and correlate with intrusive symptomatology and skin conductance measures of level of arousal (Dziobek *et al.*, 2011). Increased activation of insula in these patients has been linked to enhanced subjective negative emotional experience (Ruocco *et al.*, 2013), while PCun activation seems to play a role in BPD patients' tendency to become emotionally overinvolved in interpersonal situations (Cavanna and Trimble, 2006). Several brain imaging studies pointed out the particular importance of the amygdala for social-cognitive impairment in BPD. Amygdala hyperactivation was found during the presentation of negative scenes (Herpertz and Bertsch, 2014) as well as during the processing of neutral and emotional facial expressions (Donegan *et al.*, 2003; Minzenberg *et al.*, 2007). Finally, alterations in BPD in the mirror network system (MNS), which is crucially involved in intention recognition process, were observed in BPD. This system encompasses several areas, including inferior prefrontal gyrus, the inferior parietal lobe and the superior temporal sulcus (STS).
Overall, these altered neural responses observed in single studies are likely to reflect the result, in adult life, of a dynamic pattern of changes in cortico-limbic connectivity which mediates the relationship between the breadth of exposure to adverse childhood experiences, and the severity of adult psychopathology (Vai *et al.*, 2017). This in line with the fact that BPD patients are generally overwhelmed by automatic and affect-driven mentalizing, but fail to integrate the affective experiences with reflective and cognitive knowledge. Taken together, these data suggest psychotherapeutic interventions are most effective if they target BPD patients' mental state reasoning and cognitive ToM.

6. Mood disorders

Social cognition disabilities in major depressive disorder (MDD) are evident even in response to different types of ToM tasks (verbal/visual, cognitive/affective and reasoning/decoding) and are significantly associated to severity of depressive symptoms (Bora and Berk, 2016). Also bipolar patients, which alternate episodes of extreme euphoria, or mania, major depression, and euthymia, significantly underperformed healthy controls in ToM tasks. In this group, robust deficits were particularly reported during acute episodes of disease, while significant but modest impairments were observed in remitted and subsyndromal bipolar patients (Bora *et al.*, 2016) and in first-degree relatives of patients with bipolar disorder (Happe, 1994). In a study of Shamay-Tsoory and colleagues, euthymic BD patients showed significantly lower scores in the cognitive empathy subscale (perspective-taking), but scored significantly higher than comparison subjects on the affective empathy subscale (personal distress) (2009). These results suggest that ToM impairment may be an enduring correlate of bipolar disorder (BP), rather than a state marker of disease. This not surprising, considering that brain imaging meta-analyses and large-scale multisite studies have found that adults with BD had robust and replicable neurostructural alterations in both subcortical and cortical regions, including crucial mediating regions of ToM, such as amygdala, inferior frontal gyrus, precentral gyrus, fusiform gyrus and middle frontal cortex (Hibar *et al.*, 2018; Hibar *et al.*, 2016).
It has been suggested that social cognition impairments could be at least partly explained by cognitive deficits (Mitchell and Young, 2015). Indeed, ToM disabilities often co-exists alongside

cognitive deficits, particularly those relating to executive functions, such as inhibitory control and sustained attention (Van Rheenen *et al.*, 2014, Wolf *et al.*, 2010) and some of them correlate with, or predict, the degree of ToM impairment (Moriguchi, 2014). Longitudinal investigations are needed to assess the trajectory of socio-cognitive profile of BP across mood states and disentangle the effect of other clinical symptoms, such as cognitive functioning, from ToM ability. The presence of socio-cognitive deficits has significant clinical value, given evidence that such alterations constitute an important obstacle for social integration and predict the likelihood of future decline in social functioning (Purcell *et al.*, 2013). Moreover, monitoring of ToM impairment in euthymic states of BP might potentially prove a useful indicator of relapse (Barrera *et al.*, 2013). Altogether, these observations warrant interest in potentiating remediation program for social cognition in BD (Lahera *et al.*, 2013). And again, a clinical condition which has been associated to multiple biological disturbances, ranging from chronobiological abnormalities to brain white matter pathology (Wirz-Justice & Benedetti, 2020), involves ToM and empathy deficits which can reveal and be assessed in proper experimental settings, but which are also very closely related to the clinical phenotype of the disorder (e.g., to the wrong interpretations of others' attitudes as expression of reproach and contempt against oneself, during depression; or in term of approval and support, during mania).

As a general comment to the three clinical situations presented here, we can then conclude that independent of the specific psychobiological underpinnings and clinical manifestations of the disorders, the RDoC perspective can be useful in providing a new framework for the assessment of psychopathology. Rather obviously, this cannot be considered as the only useful perspective in approaching psychiatric conditions, but it specifically allows to study dimensions which are common to different psychiatric conditions, and to normal human behavior, thus linking experimental settings with the clinical dissection of complex psychiatric phenotypes, with the specific aim to explore their biological correlates in a neuroscience research perspective, and to (try to) provide operational definitions of manifestations at the behavioural and neuropsychological level. Confirming that this perspective cannot be limited to psychiatry, a systematic review of the literature found ToM and empathy deficits both, in other psychiatric conditions (eating disorders, obsessive-compulsive disorder, substance abuse), in neurological conditions (Alzheimer's disease and other dementias, Parkinson's disease, Huntington's disease, multiple sclerosis, amyotrophic lateral sclerosis, epilepsy), and in developmental disorders (autism, attention deficit hyperactivity disorder, intellectual disability, specific language impairment) (Cotter *et al.*, 2018).

7. Finding brain correlates or reducing?

The studies described above showed neural correlates of ToM and empathy deficits in people diagnosed with schizophrenia, BPD and mood disorders. In all cases, brain structural abnormalities were found to correlate with impairments measured at the behavioural level, suggesting a biological basis for social cognition deficits. Though the experimental samples were recruited on the basis of traditional DSM diagnoses, the research units of the studies were the more fine-grained constructs of ToM and cognitive and affective empathy, in line with the suggestions of the RdoC framework that we have illustrated in Section 2 above. In this section we briefly address a conceptual issue that has been raised towards research within this framework, namely, that it is based on a reductionist assumption, that mental disorders and symptoms are just brain disorders and dysfunctions, and or that psychology and psychiatry should be reduced to neuroscience. For example, Stanghellini *et al.* (2019) claim that psychiatry is undergoing a "new reductionistic wave". There are different versions of the objection in the recent literature, but an exemplar formulation of the problem can be found in Parnas (2014):

> The RDoC's theoretical underpinning appears to be a neurocentric "type-type" reductionism: specific chunks (types) of mental life (e.g. hallucination, anhedonia) are identical with, or nothing else than, certain specific chunks (types) of neural activity (say, a certain configuration of interactions between dysfunctional neural networks). It is hard to follow the logic of Cuthbert's assertion that the RDoC is *non*-reductionistic when he repeatedly emphasizes a "mechanistic understanding" as the RDoC's ultimate goal. "Type-type" reductionism is, of course, a legitimate theoretical position, but one that is far from being universally shared and is perhaps even obsolete. (Parnas 2014, 47)

In fact, Parnas has two targets in his critique of RdoC, the first is type-type reductionism, and the other is "mechanistic understanding". Let us see them in turn.
Type-type identity theories of mental states and brain states are metaphysical theories on the nature of reality. Parnas is right in signaling that they are "obsolete". They were thoroughly discussed until about 1980, and gradually abandoned when the "multiple realizability" objection was raised. The main idea behind multiple realizability is that in principle (and often also *de facto*) a psychological kind, i.e. pain, can be or instantiated by different kinds of physical brain states, whereas identity calls for a 1:1 relation (Fodor, 1974). Now the debate in the metaphysics of mind has shifted to more refined positions, like local reductionism or disjunctive kinds type-identity, and the multiple-realizability objection itself has been critically discussed (Clapp, 2001; Dizadji-Bahmani *et al.*, 2010; Esfeld and Sachse, 2007). The details of this debate, however, arguably exhaust the entire discipline of analytical philosophy of mind, and therefore far exceed the limits of this section (see eg. Van Riel and Van Gulick, 2019 for reference). However, let us suppose we choose one of the theses of metaphysical reduction available in the philosophical debate, the more coherent one – is the RdoC framework committed to it?
The answer is no. A research project aimed at finding correlations is not committed to specifying which of the domain of variables considered is at the more fundamental level of reality. For example, a diabetologist may study the correlations between diabetes and Socio-Economic Status and believe that socio-economic events are nothing but very complex physical events, with bosons and other particles as constituents, or either, she may believe that diabetes is just a kind of chemical state, metaphysically speaking. In either case, the metaphysics and the arguments for defending it need not be part of the research project. Coming back to RdoC, this has been expressed clearly by Cuthbert and Kozak (2013):

> Granted, as with any collection of scientists, one can readily find diverse philosophical and scientific viewpoints expressed in various statements from the NIMH. However, although a conscientious reader can detect in publications emanating from the NIMH varying language on reductionism and on the role of biology vis a` vis psychology, an essential point is that the RDoC initiative does not rely upon assumptions of eliminative reductionism, or even of biological fundamentalism (e.g., Sanislow *et al.*, 2010). In this regard, it is important to note that the RDoC initiative does not depend conceptually upon a claim of mind-brain identity. (Cuthbert and Kozak 2013)

In other words, one can surely ask whether RdoC advocates are individually committed to some version of reductionism. In fact, we know that some are, as Thomas Insel, former director of the NIMH, wrote on more than one occasion that "mental disoders are biological disorders" (eg. Insel 2013). This is surely not irrelevant as a socio-epistemological point, and we will return to this at the end of this section.
Let us go back to Parnas' second target in the passage quoted above, that RdoC is reductionist in aiming at a mechanistic explanation of mental disorders. Here we are switching from

metaphysical issues to epistemic issues. Given that RdoC is definitively committed to mechanistic explanations of disorders and symptoms, the point deserves some attention. Mechanistic explanation aims at answering the question "Why this phenomenon is happening?" by describing a mechanism whose parts and actions have the phenomenon to be explained as an output. It is widely used in biology, cognitive science and medicine, and has been examined by philosophers of science in the last 30 years (See eg. Craver and Darden, 2013). A mechanistic explanation can be a reductive explanation – for example, the illusion of a ghost entering my room at night can be explained away as soon as a mechanism is identified that involves a LSD dose, my digestive and neural systems as proper parts, and the production of a ghost-image in my visual system as an output. But mechanistic explanation *need not be* a reductive explanation. To employ a mechanistic explanation precisely with the aim of preserving the autonomy of "higher levels" is a common stance for philosophers of cognitive science, for example Bill Bechtel, who is convinced of the autonomy of psychological constructs, like working memory or visual cognition. On Bechtel's account, finding out mechanisms that explain psychological phenomena and capacities at the lower neuroscientific and possibly at the biological level, can constrain our knowledge of what happen at the psychological level: "With the advent of cognitive neuroscience, mechanistic explanations of mental phenomena have increasingly included identification of the brain parts responsible for the component operations . . . . The goal of such research is not just to learn where operations occur, but to use such knowledge to further constrain and revise proposed accounts of mechanisms." (Bechtel, 2007, p. 175)

On this view, a mechanistic explanation can be part of a pluralistic explanation, in which mechanisms at different levels constrain one another, with no assumption that parts and activities on one level be "explained away" - i.e. reduced - by reference to parts and activities on another level.

In psychiatry and philosophy of psychiatry, views of this kind are currently indicated as "explanatory pluralism". The idea is that a mental disorder or symptom, as an *explanandum*, can admit a heterogeneous *explanans*, involving mechanisms when known, but also statistical correlations when mechanisms are not available. For example, a pluralistic explanation of the social impairments of schizophrenia may feature a correlation with adverse childhood events, and the detailed mechanism of a dysfunctional brain network, as we have seen in Section 5 above. Pluralism has been defended in psychiatry by Kenneth Kendler (2012, 2019). In a recent contribution to JAMA, summarizing a meta-analysis, he writes:

> I identified 306 individual predictor variables from those papers, and the variables were widely distributed across the 12 levels. Our discipline, which includes individuals with expertise in molecular biology, neuroscience, genetics, imaging, cognition, personality, clinical and developmental psychology, epidemiology, and sociology, has provided rigorous evidence of potential and widely diverse causes of psychiatric illness. (Kendler 2019, 1086)

Cuthbert and Kozak (2013) - among the advocates of RdoC - endorsed Kendler's pluralism as a possible broad scope framework.

Another look at the constructs and dimensions of the RdoC matrix suggests a genuine interplay of "high-level" psycho-social variables and "low-level" biological and neurofisiological ones, and the basis for a pluralistic account. As said before, one of the constraints for choosing a construct orsubconstruct for the matrix was "high-level", namely, that there is sufficient evidence for its validity as a functional unity of behavior the Secondly, not all the dimensions of the matrix include verbal reports and behaviour, which are non-reductive psychological constructs. As the philosopher and psychiatrist Derek Bolton has

noted, the RdoC approach leaves open the possibility that not all the dimensions of analysis have the same weight and importance for all constructs:

> some conditions that might go into the rows of the RDoC framework will have no ticks under any boxes indicating causal processes at levels other than, for instance, the genetic or the neural, such as Huntington's disease or concussion, that is, no psychological or social factors make any difference (though they may do if the row had "adjustment to"). That is to say, reductionism might be right in some cases and in some cases it is already known to be right; in other cases, the psychosocial might be more important, account for more of the variance in incidence or outcomes, than, for instance, genetic factors. (Bolton 2013, 25)

So, is RdoC not reductionist at all, and is no "new reductionistic wave" coming at all? We have just briefly argued (following recent commentators) that the philosophical theses of metaphysical and epistemic reductionism are not part and parcel of the RdoC framework. However, RdoC, just as any other scientific endeavor, can be considered both abstractly - as a framework or "theory" with constructs and principles – and concretely, as a historically and temporally located event (in the US, beginning in 2009) with participants (the experts involved in the working groups), and connected with a funding agency such as NIMH. Considerations of context of this kind pertain neither to metaphysics nor to classical epistemology, but to the broader domain of social epistemology, which investigates the various practical and political reasons involved in a scientific project. For example, the social epistemology of DSM-5 has highlighted the influence of pharmacological companies and of patients' advocacy groups in creating or eliminating disease categories (Cooper, 2014; Solomon REF). From a socio-epistemological point of view, it matters, for example, that when the project started, Tom Insel was openly in favour of a reductionist approach, while others, like Bruce Cuthbert, now seem to be more inclined to pluralism. In fact, what some commentators have hinted is that RdoC can be reductionistic in that, once established as the main research framework for psychiatry, orienting NIMH grants and thereby journal publications, it will potentially marginalize research in areas where mechanistic explanation is not feasible, like phenomenology, psychoanalysis, psychotherapy and cultural psychiatry. For example, Herschenberg and Goldfried (2015), from the point of view of behavioural psychotherapy, worry that

> If we do focus our grants so as to be consistent with the new aims, the field runs the risk of again forsaking a more comprehensive conceptual model for a reductionistic model, this time with a focus on neural circuitry as being the ultimate method for understanding an underlying disorder. (Herschenberg and Goldfried)

Likewise, Kathrin Tabb (2017). These concerns are important and correct, in that they are grounded in facts about the context and development of RdoC. Nevertheless, they are external to the conceptual structure of the project and should better be discussed on ethical and evidence-based grounds together with governments' and scientific communities research policies.

8. Concluding remarks

In this article we provided an illustration of research on mental disorders within the RdoC paradigm. Our case study has been impairments in the social domain, in particular, difficulties in Theory of Mind and empathy. Such impairments can be studied across traditional DSM categories, as they are present in Schizophrenia, Addiction Disorders and Mood Disorders populations. We showed how evidence on relevant brain anomalies is accumulating

and paving the way for more accurate explanations of what is it to have difficulties in understanding others. In the last section, we zoomed out and considered this kind of research vis-à-vis the objection of being reductionistic that is, in favoring mechanistic accounts of dysfunctionalities. We argued on philosophical grounds, that metaphysical reductionism and explanatory reductionism are not conceptually entailed by the RdoC framework. However, the possibility of keeping psychiatry genuinely pluralistic is a question of research policy.

REFERENCES

Baron-Cohen S., Leslie A. M. & Frith, U. (1985). Does the autistic child have a "theory of mind". *Cognition*, *21*(1), pp. 37-46;
Casey B. J., Craddock N., Cuthbert B. N., Hyman S. E., Lee F. S. & Ressler K. J. (2013). DSM-5 and RDoC: progress in psychiatry research?. *Nature Reviews Neuroscience*, *14*(11), pp. 810-814;
Charney D. S., Buxbaum J. D., Sklar P. & Nestler E. J. (Eds.). (2013). *Neurobiology of mental illness*. Oxford University Press;
Cotter J., Granger K., Backx R., Hobbs M., Looi C. Y. & Barnett J. H. (2018). Social cognitive dysfunction as a clinical marker: a systematic review of meta-analyses across 30 clinical conditions. *Neuroscience & Biobehavioral Reviews*, pp. *84*, 92-99;
Cuthbert B. N. & Kozak M. J. (2013). Constructing constructs for psychopathology: the NIMH research domain criteria. *Journal of abnormal psychology*, *122*(3), pp. 928-937;
Cuthbert B. N. & Insel T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RdoC. *BMC medicine*, *11*(1), p. 126;
Cuthbert B. N. (2014). The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, *13*(1), pp. 28-35;
Decety J., & Jackson P. L. (2004). The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, *3*(2), pp. 71-100;
de Waal F. B. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual review of psychology*, pp. *59*, p. 279;
Faucher L. & Goyer S. (2015). RDoC: Thinking outside the DSM box without falling into a reductionist trap. In *The DSM-5 in perspective* (pp. 199-224). Springer, Dordrecht;
Gallese V. (2007). Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), pp. 659-669;
Harrison L. A., Kats, A. Williams M. E. & Aziz-Zadeh L. (2019). The importance of sensory processing in mental health: A proposed addition to the Research Domain Criteria (RDoC) and suggestions for RDoC 2.0. *Frontiers in psychology*, *10*, p. 103;
Hoffman G.A. and Zachar P. (2017) RDoC's Metaphysical Assumptions: Problems and Promises." In *Extraordinary Science: Responding to the Current Crisis in Psychiatric Research*, ed. Jeffrey Poland, and Şerife Tekin, pp. 59-86. Cambridge: MIT Press;
Hyman S. E. (2007). Can neuroscience be integrated into the DSM-V?. *Nature Reviews Neuroscience*, *8*(9), 725-732;
Hyman S. E. (2010). The diagnosis of mental disorders: the problem of reification. *Annual review of clinical psychology*, *6*, pp. 155-179; National Institute of Mental Health (NIMH)
Insel T., Cuthbert B., Garvey M., Heinssen R., Pine D. S., Quinn K. ... & Wang P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders;
Lake J., Yee C. & Miller G. Misunderstanding RDoC. *Zeitschrift für Psychologie*, *225*(3), pp. 170-174;
London E. B. (2014). Categorical diagnosis: a fatal flaw for autism research?. *Trends in neurosciences*, *37*(12), pp. 683-686;

Koudys J. W., Traynor J. M., Rodrigo A. H., Carcone D. & Ruocco A. C. (2019). The NIMH Research Domain Criteria (RDoC) initiative and its implications for research on personality disorder. *Current psychiatry reports*, *21*(6), p. 37.
Lilienfeld S. O. & Treadway M. T. (2016). Clashing diagnostic approaches: DSM-ICD versus RDoC. *Annual review of clinical psychology*, *12*, pp. 435-463;
Mitchell R. L. & Phillips L. H. (2015). The overlapping relationship between emotion perception and theory of mind. *Neuropsychologia*, *70*, pp. 1-10;
Molenberghs P., Johnson H., Henry J. D. & Mattingley J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, *65*, pp. 276-291;
Morris S. E. & Cuthbert B. N. (2012). Research Domain Criteria: cognitive systems, neural circuits, and dimensions of behavior. *Dialogues in clinical neuroscience*, *14*(1), p. 29;
Nesse R. M. & Stein D. J. (2012). Towards a genuinely medical model for psychiatric nosology. *BMC medicine*, *10*(1), p. 5;
National Institute of Mental Health (NIMH) (2019) https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/about-rdoc.shtml;
Premack D. & Woodruff G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and brain sciences*, *1*(4), pp. 515-526;
Stanislow C., Morris S., Pacheco J., Cuthbert B. (2020). The National Institute of Mental Health Research Domain Criteria: an alternative framework to guide psychopathology research. In Geddes J. R. & Andreasen N. C. (2020). *New Oxford textbook of psychiatry*. Oxford University Press, USA, pp. 62-73;
Stueber, Karsten, "Empathy", *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/empathy/>;
Tabb, K. (2015). Psychiatric progress and the assumption of diagnostic discrimination. *Philosophy of Science*, *82*(5), pp. 1047-1058;
Tomasello M. & Moll H. (2013). Why don't apes understand false beliefs? In MR Banaji & SA Gelman (Eds.), Navigating the social world: What infants, children, and other species can teach us (pp. 81-88). Oxford [ua]: Oxford Univ;
Tsou J. Y. (2015). DSM-5 and psychiatry's second revolution: Descriptive vs. theoretical approaches to psychiatric classification. In *The DSM-5 in perspective* (pp. 43-62). Springer, Dordrecht;
Walther S., Bernard J. A., Mittal V. A. & Shankman S. A. (2019). The utility of an RDoC motor domain to understand psychomotor symptoms in depression. *Psychological medicine*, *49*(2), pp. 212-216;
Wellman H. M. & Liu D. (2004). Scaling of theory-of-mind tasks. *Child development*, *75*(2), pp. 523-541;
World Health Organization (WHO) (2018). *International classification of diseases for mortality and morbidity statistics* (11th Revision). Retrieved from https://icd.who.int/browse11/l-m/en;
Yee C. M., Javitt D. C. & Miller G. A. (2015). Replacing DSM categorical analyses with dimensional analyses in psychiatry research: the research domain criteria initiative. *JAMA psychiatry*, *72*(12), pp. 1159-1160;
Zachar P., Stoyanov D. S., Aragona M. & Jablensky A. (Eds.). (2014). *Alternative perspectives on psychiatric validation: DSM, ICD, RDoC, and beyond*. OUP Oxford;
Adrian J. E., Clemente R. A., Villanueva L. & Rieffe C. (2005). Parent-child picture-book reading, mothers' mental state language and children's theory of mind. *J Child Lang* 32, pp. 673-86;
Amodio D. M. & Frith C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci* 7, pp. 268-77;
Arntz A., Bernstein D., Oorschot M. & Schobre P. (2009). Theory of mind in borderline and cluster-C personality disorder. *J Nerv Ment Dis* 197, pp. 801-7.

Backasch B., Straube B., Pyka M., Klohn-Saghatolislam F., Muller M. J., Kircher T. T. & Leube D. T. (2013). Hyperintentionality during automatic perception of naturalistic cooperative behavior in patients with schizophrenia. *Soc Neurosci* 8, pp. 489-504;
Baron-Cohen S., O'Riordan M., Stone V., Jones R. & Plaisted K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *J Autism Dev Disord* 29, pp. 407-18;
Baron-Cohen S., Wheelwright S., Hill J., Raste Y. & Plumb I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J Child Psychol Psychiatry* 42, pp. 241-51;
Barrera A., Vazquez G., Tannenhaus L., Lolich M. & Herbst L. (2013). Theory of mind and functionality in bipolar patients with symptomatic remission. *Rev Psiquiatr Salud Ment* 6, pp. 67-74;
Benedetti F., Bernasconi A., Bosia M., Cavallaro R., Dallaspezia S., Falini A., Poletti S., Radaelli D., Riccaboni R., Scotti G. & Smeraldi E. (2009). Functional and structural brain correlates of theory of mind and empathy deficits in schizophrenia. *Schizophr Res* 114, pp. 154-60;
Bora E. (2009). [Theory of mind in schizophrenia spectrum disorders]. *Turk Psikiyatri Derg* 20, pp. 269-81;
Bora, E., Bartholomeusz, C. & Pantelis, C. (2016). Meta-analysis of Theory of Mind (ToM) impairment in bipolar disorder. *Psychol Med* 46, pp. 253-64;
Bora E. & Berk M. (2016). Theory of mind in major depressive disorder: A meta-analysis. *J Affect Disord* 191, pp. 49-55;
Bora E. & Pantelis C. (2016). Social cognition in schizophrenia in comparison to bipolar disorder: A meta-analysis. *Schizophr Res* 175, pp. 72-78;
Bosia M., Riccaboni R. & Poletti S. (2012). Neurofunctional correlates of theory of mind deficits in schizophrenia. *Curr Top Med Chem* 12, pp. 2284-302;
Brothers L. & Ring B. (1992). A neuroethological framework for the representation of minds. *J Cogn Neurosci* 4, pp. 107-18;
Brune M., Lissek S., Fuchs N., Witthaus H., Peters S., Nicolas V., Juckel G. & Tegenthoff M. (2008). An fMRI study of theory of mind in schizophrenic patients with "passivity" symptoms. *Neuropsychologia* 46, pp. 1992-2001;
Brunet E., Sarfati Y., Hardy-Bayle M. C. & Decety J. (2003). Abnormalities of brain function during a nonverbal theory of mind task in schizophrenia. *Neuropsychologia* 41, pp. 1574-82;
Callicott J. H., Bertolino A., Mattay V. S., Langheim F. J., Duyn J., Coppola R., Goldberg T. E. & Weinberger D. R. (2000). Physiological dysfunction of the dorsolateral prefrontal cortex in schizophrenia revisited. *Cereb Cortex* 10, pp. 1078-92;
Cavanna A. E. & Trimble M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain* 129, pp. 564-83;
Ciaramidaro A., Bolte, S., Schlitt S., Hainz D., Poustka F., Weber B., Bara B. G., Freitag C. & Walter H. (2015). Schizophrenia and autism as contrasting minds: neural evidence for the hypo-hyper-intentionality hypothesis. *Schizophr Bull* 41, pp. 171-9;
Corcoran R., Cahill C. & Frith C. D. (1997). The appreciation of visual jokes in people with schizophrenia: a study of 'mentalizing' ability. *Schizophr Res* 24, pp. 319-27;
Cotter J., Granger K., Backx R., Hobbs M., Looi C. Y. & Barnett J. H. (2018). Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neurosci Biobehav Rev* 84, pp. 92-99;
Cuthbert B. N. & Insel T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med* 11, p. 126;

Dapretto M., Davies M. S., Pfeifer J. H., Scott A. A., Sigman M., Bookheimer S. Y. & Iacoboni M. (2006). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nat Neurosci* 9, pp. 28-30;
Davidson C. A., Piskulic D., Addington J., Cadenhead K. S., Cannon T. D., Cornblatt B. A., McGlashan T. H., Perkins D. O., Seidman L. J., Tsuang M. T., Walker E. F., Bearden C. E., Mathalon D. H., Woods S. W. & Johannesen J. K. (2018). Age-related trajectories of social cognition in youth at clinical high risk for psychosis: An exploratory study. *Schizophr Res* 201, pp. 130-136;
Domes G., Schulze L. & Herpertz S. C. (2009). Emotion recognition in borderline personality disorder-a review of the literature. *J Pers Disord* 23, pp. 6-19;
Donegan N. H., Sanislow C. A., Blumberg H. P., Fulbright R. K., Lacadie C., Skudlarski P., Gore J. C., Olson I. R., McGlashan T. H. & Wexler B. E. (2003). Amygdala hyperreactivity in borderline personality disorder: implications for emotional dysregulation. *Biol Psychiatry* 54, pp. 1284-93;
Dziobek I., Preissler S., Grozdanovic Z., Heuser I., Heekeren H. R. & Roepke S. (2011). Neuronal correlates of altered empathy and social cognition in borderline personality disorder. *Neuroimage* 57, pp. 539-48;
Farrow T. F., Whitford T. J., Williams L. M., Gomes L. & Harris A. W. (2005). Diagnosis-related regional gray matter loss over two years in first episode schizophrenia and bipolar disorder. *Biol Psychiatry* 58, pp. 713-23;
Fertuck E. A., Jekal A., Song I., Wyman B., Morris M. C., Wilson S. T., Brodsky B. S. & Stanley B. (2009). Enhanced 'Reading the Mind in the Eyes' in borderline personality disorder compared to healthy controls. *Psychol Med* 39, pp. 1979-88;
Fett A. K., Viechtbauer W., Dominguez M. D., Penn D. L., van Os J. & Krabbendam L. (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: a meta-analysis. *Neurosci Biobehav Rev* 35, pp. 573-88;
Fonagy P. & Target M. (1996). Playing with reality: I. Theory of mind and the normal development of psychic reality. *Int J Psychoanal* 77 (Pt 2), pp. 217-33;
Franzen N., Hagenhoff M., Baer N., Schmidt A., Mier D., Sammer G., Gallhofer B., Kirsch, P. & Lis S. (2011). Superior 'theory of mind' in borderline personality disorder: an analysis of interaction behavior in a virtual trust game. *Psychiatry Res* 187, pp. 224-33;
Frick C., Lang S., Kotchoubey B., Sieswerda S., Dinu-Biringer R., Berger M., Veser S., Essig M. & Barnow S. (2012). Hypersensitivity in borderline personality disorder during mindreading. *PLoS One* 7, e41650.
Frith C. D. & Frith U. (2007). Social cognition in humans. *Curr Biol* 17, R724-32.
Ghiassi V., Dimaggio G. & Brune M. (2010). Dysfunctions in understanding other minds in borderline personality disorder: a study using cartoon picture stories. *Psychother Res* 20, pp. 657-67;
Greig T. C., Bryson G. J. & Bell M. D. (2004). Theory of mind performance in schizophrenia: diagnostic, symptom, and neuropsychological correlates. *J Nerv Ment Dis* 192, pp. 12-8;
Gross & Harris (1988). False Beliefs About Emotion: Children's Understanding of Misleading Emotional Displays. *International Journal of Behavioral Development*, pp. 475-488;
Happe F. G. (1994). An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *J Autism Dev Disord* 24, pp. 129-54;
Herold R., Tenyi T., Lenard K. & Trixler M. (2002). Theory of mind deficit in people with schizophrenia during remission. *Psychol Med* 32, pp. 1125-9;
Herpertz S. C. & Bertsch K. (2014). The social-cognitive basis of personality disorders. *Curr Opin Psychiatry* 27, pp. 73-7;

Hershenberg R. & Goldfried M. R. (2015). Implications of RDoC for the research and practice of psychotherapy. *Behavior Therapy*, 46(2), pp. 156-165.
Hibar D. P., Westlye L. T., Doan N. T., Jahanshad N., Cheung J. W., Ching C. R. K., Versace A., Bilderbeck A. C., Uhlmann A., Mwangi B., Kramer B., Overs B., Hartberg C. B., Abe C., Dima D., Grotegerd D., Sprooten E., Boen E., Jimenez E., Howells F. M., Delvecchio G., Temmingh H., Starke J., Almeida J. R. C., Goikolea J. M., Houenou J., Beard L. M., Rauer L., Abramovic L., Bonnin M., Ponteduro M. F., Keil M., Rive M. M., Yao N., Yalin N., Najt P., Rosa P. G., Redlich R., Trost S., Hagenaars S., Fears S. C., Alonso-Lana S., van Erp T. G. M., Nickson T., Chaim-Avancini T. M., Meier T. B., Elvsashagen T., Haukvik U. K., Lee W. H., Schene A. H., Lloyd A. J., Young A. H., Nugent A., Dale A. M., Pfennig A., McIntosh A. M., Lafer B., Baune B. T., Ekman C. J., Zarate C. A., Bearden C. E., Henry C., Simhandl C., McDonald C., Bourne C., Stein D. J., Wolf D. H., Cannon D. M., Glahn D. C., Veltman D. J., Pomarol-Clotet E., Vieta, E., Canales-Rodriguez E. J., Nery F. G., Duran F. L. S., Busatto G. F., Roberts G., Pearlson G. D., Goodwin G. M., Kugel H., Whalley H. C., Ruhe H. G., Soares J. C., Fullerton J. M., Rybakowski J. K., Savitz J., Chaim K. T., Fatjo-Vilas M., Soeiro-de-Souza M. G., Boks M. P., Zanetti M. V., Otaduy M. C. G., Schaufelberger M. S., Alda M., Ingvar M., Phillips M. L., Kempton M. J., Bauer M., Landen M., Lawrence N. S., van Haren N. E. M., Horn N. R., Freimer N. B., Gruber O., Schofield P. R., Mitchell P. B., Kahn R. S., Lenroot R., Machado-Vieira R., Ophoff R. A., Sarro S., Frangou S., Satterthwaite T. D., Hajek T., Dannlowski U., Malt U. F., Arolt V., Gattaz W. F., Drevets W. C., Caseras X., Agartz I., Thompson P. M. & Andreassen O. A. (2018). Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. *Mol Psychiatry* 23, pp. 932-942;
Hibar D. P., Westlye L. T., van Erp T. G., Rasmussen J., Leonardo C. D., Faskowitz J., Haukvik U. K., Hartberg C. B., Doan N. T., Agartz I., Dale A. M., Gruber O., Kramer B., Trost S., Liberg B., Abe C., Ekman C. J., Ingvar M., Landen M., Fears S. C., Freimer N. B., Bearden C. E., Costa Rica/ Colombia Consortium for Genetic Investigation of Bipolar, Sprooten E., Glahn D. C., Pearlson G. D., Emsell L., Kenney J., Scanlon C., McDonald C., Cannon D. M., Almeida J., Versace A., Caseras X., Lawrence N. S., Phillips M. L., Dima D., Delvecchio G., Frangou S., Satterthwaite T. D., Wolf D., Houenou J., Henry C., Malt U. F., Boen E., Elvsashagen T., Young A. H., Lloyd A. J., Goodwin G. M., Mackay C. E., Bourne C., Bilderbeck A., Abramovic L., Boks M. P., van Haren N. E., Ophoff R. A., Kahn R. S., Bauer M., Pfennig A., Alda M., Hajek T., Mwangi B., Soares J. C., Nickson T., Dimitrova R., Sussmann J. E., Hagenaars S., Whalley H. C., McIntosh A. M., Thompson P. M. & Andreassen O. A. (2016). Subcortical volumetric abnormalities in bipolar disorder. *Mol Psychiatry* 21, pp. 1710-1716;
Horan W. P., Nuechterlein K. H., Wynn J. K., Lee J., Castelli F. & Green M. F. (2009). Disturbances in the spontaneous attribution of social meaning in schizophrenia. *Psychol Med* 39, pp. 635-43;
Hyman S. E. (2007). Can neuroscience be integrated into the DSM-V? *Nat Rev Neurosci* 8, pp. 725-32;
Hyman S. E. (2010). The diagnosis of mental disorders: the problem of reification. *Annu Rev Clin Psychol* 6, pp. 155-79;
Inoue Y., Yamada K., Hirano M., Shinohara M., Tamaoki T., Iguchi H., Tonooka Y. & Kanba S. (2006). Impairment of theory of mind in patients in remission following first episode of schizophrenia. *Eur Arch Psychiatry Clin Neurosci* p. 256, pp. 326-8;
Insel T., Cuthbert B., Garvey M., Heinssen R., Pine D. S., Quinn K., Sanislow C. & Wang P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry* p. 167, pp. 748-51;
Kelemen O., Erdelyi R., Pataki I., Benedek G., Janka Z. & Keri S. (2005). Theory of mind and motion perception in schizophrenia. *Neuropsychology* 19, pp. 494-500;
Kendell R. & Jablensky A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *Am J Psychiatry* 160, pp. 4-12;

Kern R. S., Green M. F., Fiske A. P., Kee K. S., Lee J., Sergi M. J., Horan W. P., Subotnik K. L., Sugar C. A. & Nuechterlein K. H. (2009). Theory of mind deficits for processing counterfactual information in persons with chronic schizophrenia. *Psychol Med* 39, pp. 645-54;
Kington J. M., Jones L. A., Watt A. A., Hopkin E. J. & Williams J. (2000). Impaired eye expression recognition in schizophrenia. *J Psychiatr Res* 34, pp. 341-7;
Kohler C. G., Hoffman L. J., Eastman L. B., Healey K. & Moberg P. J. (2011). Facial emotion perception in depression and bipolar disorder: a quantitative review. *Psychiatry Res* 188, p. 3039;
Kohler C. G., Walker J. B., Martin E. A., Healey K. M. & Moberg P. J. (2010). Facial emotion perception in schizophrenia: a meta-analytic review. *Schizophr Bull* 36, pp. 1009-19;
Kronbichler L., Tschernegg M., Martin A. I., Schurz M. & Kronbichler M. (2017). Abnormal Brain Activation During Theory of Mind Tasks in Schizophrenia: A Meta-Analysis. *Schizophr Bull* 43, pp. 1240-1250;
Lahera G., Benito A. Montes J. M., Fernandez-Liria A., Olbert C. M. & Penn D. L. (2013). Social cognition and interaction training (SCIT) for outpatients with bipolar disorder. *J Affect Disord* 146, pp. 132-6;
Lazarus S. A., Cheavens J. S., Festa F. & Zachary Rosenthal M. (2014). Interpersonal functioning in borderline personality disorder: a systematic review of behavioral and laboratory-based assessments. *Clin Psychol Rev* 34, pp. 193-205;
Leppanen J., Sedgewick F., Treasure J. & Tchanturia K. (2018). Differences in the Theory of Mind profiles of patients with anorexia nervosa and individuals on the autism spectrum: A meta-analytic review. *Neurosci Biobehav Rev* 90, pp. 146-163;
Leslie A. M. & Keeble S. (1987). Do six-month-old infants perceive causality? *Cognition* 25, pp. 265-88;
Malhi G. S., Lagopoulos J., Das P., Moss K., Berk M. & Coulston C. M. (2008). A functional MRI study of Theory of Mind in euthymic bipolar disorder patients. *Bipolar Disord* 10, pp. 943-56;
Mar R. A. (2011). The neural bases of social cognition and story comprehension. *Annu Rev Psychol* 62, pp. 103-34;
Martin A. K., Robinson G., Dzafic I., Reutens D. & Mowry B. (2014). Theory of mind and the social brain: implications for understanding the genetic basis of schizophrenia. *Genes Brain Behav* 13, pp. 104-17;
Meristo M., Falkman K. W., Hjelmquist E., Tedoldi M., Surian L. & Siegal M. (2007). Language access and theory of mind reasoning: evidence from deaf children in bilingual and oralist environments. *Dev Psychol* 43, pp. 1156-69;
Mier D., Lis S., Esslinger C., Sauer C., Hagenhoff M., Ulferts J., Gallhofer B. & Kirsch P. (2013). Neuronal correlates of social cognition in borderline personality disorder. *Soc Cogn Affect Neurosci* 8, pp. 531-7;
Milders M., Ietswaart M., Crawford J. R. & Currie D. (2006). Impairments in theory of mind shortly after traumatic brain injury and at 1-year follow-up. *Neuropsychology* 20, pp. 400-408;
Milligan K., Astington J. W. & Dack L. A. (2007). Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child Dev* 78, pp. 622-46;
Minzenberg M. J., Fan J., New A. S., Tang C. Y. & Siever L. J. (2007). Fronto-limbic dysfunction in response to facial emotion in borderline personality disorder: an event-related fMRI study. *Psychiatry Res* 155, pp. 231-43;
Mitchell A. E., Dickens G. L. & Picchioni M. M. (2014). Facial emotion processing in borderline personality disorder: a systematic review and meta-analysis. *Neuropsychol Rev* 24, pp. 166-84;
Mitchell R. L. & Young A. H. (2015). Theory of Mind in Bipolar Disorder, with Comparison to the Impairments Observed in Schizophrenia. *Front Psychiatry* 6, pp. 188;

Mo S., Su Y., Chan R. C. & Liu J. (2008). Comprehension of metaphor and irony in schizophrenia during remission: the role of theory of mind and IQ. *Psychiatry Res* 157, pp. 21-9;
Moriguchi Y. (2014). The early development of executive function and its relation to social interaction: a brief review. *Front Psychol* 5, p. 388;
Moses S. N., Ostreicher M. L. & Ryan J. D. (2010). Relational framework improves transitive inference across age groups. *Psychol Res* 74, pp. 207-18;
Murphy D. (2006). Theory of mind in Asperger's syndrome, schizophrenia and personality disordered forensic patients. *Cogn Neuropsychiatry* 11, pp. 99-111;
Nemeth N., Matrai P., Hegyi P., Czeh B., Czopf L., Hussain A., Pammer J., Szabo I., Solymar M., Kiss L., Hartmann P., Szilagyi A. L., Kiss Z. & Simon M. (2018). Theory of mind disturbances in borderline personality disorder: A meta-analysis. *Psychiatry Res* 270, pp. 143-153;
Perner J. & Davies G. (1991). Understanding the mind as an active information processor: do young children have a "copy theory of mind"? *Cognition* 39, pp. 51-69;
Perner J., Mauer M. C. & Hildenbrand M. (2011). Identity: key to children's understanding of belief. *Science* 333, pp. 474-7;
Perner J. & Winner H. (1985). John thinks that Mary thinks that..." attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology* 39;
Povinelli D. J. (1993). Reconstructing the evolution of mind. *Am Psychol* 48, 493-509;
Purcell A. L., Phillips M. & Gruber J. (2013). In your eyes: does theory of mind predict impaired life functioning in bipolar disorder? *J Affect Disord* 151, pp. 1113-9;
Ruocco A. C., Amirthavasagam S., Choi-Kain L. W. & McMain S. F. (2013). Neural correlates of negative emotionality in borderline personality disorder: an activation-likelihood-estimation meta-analysis. *Biol Psychiatry* 73, pp. 153-60;
Samson D., Apperly I. A. & Humphreys G. W. (2007). Error analyses reveal contrasting deficits in "theory of mind": neuropsychological evidence from a 3-option false belief task. *Neuropsychologia* 45, pp. 2561-9;
Sanislow C. A., Pine D. S., Quinn K. J., Kozak M. J., Garvey M. A., Heinssen R. K., Wang P. S. & Cuthbert B. N. (2010). Developing constructs for psychopathology research: research domain criteria. *J Abnorm Psychol* 119, pp. 631-9.
Saxe R. & Wexler A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia* 43, pp. 1391-9;
Schilling L., Wingenfeld K., Lowe B., Moritz S., Terfehr K., Kother U. & Spitzer C. (2012). Normal mind-reading capacity but higher response confidence in borderline personality disorder patients. *Psychiatry Clin Neurosci* 66, pp. 322-7;
Schurz M., Radua J., Aichhorn M., Richlan F. & Perner J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neurosci Biobehav Rev* 42, pp. 9-34;
Scott L. N., Levy K. N., Adams R. B., Jr. & Stevenson M. T. (2011). Mental state decoding abilities in young adults with borderline personality disorder traits. *Personal Disord* 2, pp. 98-112;
Shamay-Tsoory S., Harari H., Szepsenwol O. & Levkovitz Y. (2009). Neuropsychological evidence of impaired cognitive empathy in euthymic bipolar disorder. *J Neuropsychiatry Clin Neurosci* 21, pp. 59-67;
Shamay-Tsoory S. G. (2011). The neural bases for empathy. *Neuroscientist* 17, pp. 18-24;
Shamay-Tsoory S. G., Shur S., Barcai-Goodman L., Medlovich S., Harari H. & Levkovitz Y. (2007). Dissociation of cognitive from affective components of theory of mind in schizophrenia. *Psychiatry Res* 149, pp. 11-23;
Sullivan Winner & Hopfield (1995). How children tell a lie from a joke: The role of second-order mental state attributions. *Developmental Psychology*, pp. 191–204;

Takahashi T., Wood S. J., Yung A. R., Soulsby B., McGorry P. D., Suzuki M., Kawasaki Y., Phillips L. J., Velakoulis D. & Pantelis C. (2009). Progressive gray matter reduction of the superior temporal gyrus during transition to psychosis. *Arch Gen Psychiatry* 66, pp. 366-76;
Thye M. D., Murdaugh D. L. & Kana R. K. (2018). Brain Mechanisms Underlying Reading the Mind from Eyes, Voice, and Actions. *Neuroscience* 374, pp. 172-186;
Unoka Z., Fogd D., Fuzy M. & Csukly G. (2011). Misreading the facial signs: specific impairments and error patterns in recognition of facial emotions with negative valence in borderline personality disorder. *Psychiatry Res* 189, pp. 419-25;
Van Overwalle F. & Baetens K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage* 48, pp. 564-84;
Van Rheenen T. E., Meyer D. & Rossell S. L. (2014). Pathways between neurocognition, social cognition and emotion regulation in bipolar disorder. *Acta Psychiatr Scand* 130, pp. 397-405;
Vogeley K., Bussfeld P., Newen A., Herrmann S., Happe F., Falkai P., Maier W., Shah N. J., Fink G. R. & Zilles K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage* 14, pp. 170-81;
Walter H., Ciaramidaro A., Adenzato M., Vasic N., Ardito R. B., Erk S. & Bara B. G. (2009). Dysfunction of the social brain in schizophrenia is modulated by intention type: an fMRI study. *Soc Cogn Affect Neurosci* 4, pp. 166-76;
Wolf F., Brune M. & Assion H. J. (2010). Theory of mind and neurocognitive functioning in patients with bipolar disorder. *Bipolar Disord* 12, pp. 657-66.