Dr. Krystyna Bielecka

Psychopathology From an Anti-Representational Perspective. A Critical Analysis

1. Research project objectives

The current project concerns the understanding of various kinds of psychopathology from an anti-representational perspective. In anti-representational accounts, it is claimed that one need not appeal to semantic properties of mental states to explain and describe mental disorders. The standard explanation of depression can be interpreted anti-representionally, because, as the proponents of this position indicate, depression is not "about something", i.e., it has no content or object (such a position is defended by, e.g., John Searle (1983)). Representationalists, unlike anti-representationalists, may argue that, say, depression is *about* the whole world to which the person suffering from depression has a negative attitude (Jean-Paul Sartre (1948) and Tim Crane (1998) argued for this point). Representations are understood here as having referents (if any), contents (satisfaction conditions) and vehicles (physical bearers that realize their representational states). Satisfaction conditions indicate when the representation is accurate (e.g., satisfied by a referent in the case of simple entity representations or true in the case of propositional contents). In particular, representationalists appeal to mental representations, which are mental states or processes that have the features of representations.

The objective of the current project is to determine whether any specific kind of psychopathology is representational or involves representational phenomena and if so, why. To achieve this end, the research team will refer to the preliminary work on the notion of mental representation in the philosophy of mind and cognitive science that was the topic of the doctoral dissertation of the Principal Investigator (PI). The PI has proposed a teleosemantic account of mental representation whose crucial requirement is for representational mechanisms to be able to detect inconsistencies in order to discover representational error. The model will be assumed as a working theoretical assumption of the project. Making one assumption at the outset is required simply because there are multiple mutually incompatible models of mental representation, and some of them turn out to be extremely liberal, rendering the notion of representation trivial as a result (Ramsey 2007). However, in the event that the assumption turns out to be too strict or too imprecise, the model will be modified. As a result, the research team will reconsider the list of desiderata for a satisfactory account of mental representation already compiled by the PI to make them more precise and possibly to include some further desiderata that fit the psychopathology research better. Furthermore, the team will reassess the coherence-based model of detectable representational error in an effort to fit more complex kinds of psychopathology.

The research team will investigate some cases of psychopathology whose representational character is usually taken for granted (such as delusions), as well as those whose status is unclear (confabulations) or even those that are commonly understood anti-representationally (obsessive-compulsive disorders in animals, see also the preliminary work in (Bielecka i Marcinów, submitted)) in order to work out a representational account of psychopathology and analyze in depth the arguments made by proponents of anti-representationalism.

In addition, a taxonomy of psychopathology will be introduced to distinguish between those that are representational and those that are anti-representational according to the model assumed in the project. To do this, selected cases of psychopathology will be used as test cases for the model, and if necessary, the model will be refined. Some simple kinds of psychopathology will probably fit the already sketched minimal model of mental representation better, and the more complex cases of mental disorder will most probably require some changes and additions to the model. It is possible that the current coherence-based model does not fit some kinds of psychopathology at all; some of its features may need to be changed or a case under investigation may not be representational at all, contrary to the claims of extreme representationalists such as Tim Crane.

The initial research hypotheses are as follows:
(H1) at least some kinds of psychopathology need to be understood anti-representationally (*contra* Crane and other extreme representationalists), and
(H2) at least some kinds of psychopathology need to be understood representationally (*contra* extreme anti-representationalists);
(H3) a coherence-based model can be applied not only to simple but also to complex cases of psychopathology (that involve multiple disorders), and
(H4) psychopathology of a representational nature must have at least minimal content (have satisfaction conditions, contra Hutto and Myin 2013);

(H5) there are many types of representational errors that can be found in different types of psychopathology, some of which are related to the operation of the representation producer, and some to the operation of the consumer.
The producer of the representation is a subsystem responsible for creating representational states (similar to a message sender in simple models of communication), and the consumer – a subsystem that responds to such states, just like the message receiver (cf. Millikan 1984).

To validate these hypotheses, one needs to investigate particular cases as studied in the literature and clinical practice. The investigation of the cases may result in some changes in the account of representation, either a list of desiderata or the initial coherence-based model. It will also lead to a more general assessment of the utility of methodological anti-representationalism in the field of psychopathology.

2. **Significance of the project** *2.1. State of the art*

Mental representations are *intentional* mental states or processes, as they are *about* something (Brentano 1900; for a review of positions, see also Haugeland 1990). They have semantic properties, such as content (usually equated with satisfaction conditions), reference (an object, a feature of an object, state of affairs, *etc.*) and a vehicle (e.g., a proposition in the language of thought or a neuronal process of the central nervous system, *etc.*). Thus philosophically characterized mental representations play an important role in folk psychology because they can be used in everyday explanations and descriptions of human action.

In analytic philosophy, mental representations have usually been understood as propositional attitudes, or various attitudes that people have towards propositions (Fodor 1978; Matthews 2007). Thoughts, beliefs, desires, perceptions or imaginings can be all accounted for in terms of propositional attitudes. It has been claimed that propositional attitudes are at the core of folk psychology, and some have defended the view that psychology as a science also rests on propositional attitudes understood as relations between organisms and internal representations (Fodor 1975, 1978). Such mental representations are characterized in a Cartesian manner as inner (but not necessarily conscious) states. They are systematically organized, functional internal states and have a linguistic form in a language of thought. The best example of this view is Fodor's Representational Theory of Mind (Fodor 1975), later relaxed in order to include mental images as well (Fodor 2008).

Others have argued that the relationship between folk psychological entities such as beliefs and internal representational states may be much more complex, and that there may be no straightforward correspondence between them (Dennett 1987). They have stressed that in cognitive sciences, mental representations are usually characterized as sub-personal states that have biological and functional roles or a computational role in a cognitive system (Dennett 1969; Drayson 2014). For this reason, some defended the view that propositional attitudes can be characterized best in terms of behavioral dispositions (Matthews 2007; Schwitzgebel 2002). These are not necessarily internal representations at all.

There is an ongoing debate over the possibility of further explaining the nature of folk psychological representations and their underlying cognitive mechanisms while maintaining their philosophical features, especially their intentional character. At first glance, philosophical and scientific views on mental representations are quite different. While few philosophers have denied that there are representations in general – the existence of sentences printed on paper or stop signs is usually taken for granted even by staunch opponents of representationalism – some claim that folk psychology need not presuppose internal representation (Gallagher and Zahavi 2008; Hutto and Myin 2013).

Mental representations are usually thought to be distinct from other cognitive states, processes, or properties, such as phenomenal states (Chalmers 1996). Some however have claimed that there is an inextricable link between (phenomenal) consciousness and intentionality (Mendelovici and Bourget 2014; Searle 1983). The latter claim that for this reason, intentional states cannot be reduced to the physical, at least not straightforwardly: they are reducible to phenomenal consciousness but phenomenal consciousness may not be reducible to the physical. However, the purported detailed elucidation of how consciousness gives rise to intentionality does not seem to be forthcoming, and it may be questioned that the view is compatible with naturalism, which is a common assumption in the modern philosophy of mind. For example, John Searle has claimed that consciousness is subjective, and at the same time *caused* by the neurobiological mechanisms, which implies, in his view, the possibility of causal reduction without ontological reduction (Searle 2002). It is however unclear how exactly neurobiological mechanisms give rise to satisfaction conditions of thoughts or beliefs in Searle's view, and it remains quite dubious that neurobiology is even in the business of accounting for satisfaction conditions (it may be in the purview of highly theoretical cognitive neuroscience but not neurobiology, as Searle boldly states).

Those who try to account for intentionality (and content) in non-intentional terms are usually functionalists, because functionalism is supposed to elucidate the causal or inferential role of representational states in a cognitive system. In contemporary functionalism, there are two views on intentionality: internalism and externalism. Internalism is the view that content supervenes on the microstructure of the cognitive system (Fodor 1980; Poczobut 2009; Segal 2000), and is more likely to appeal to the notion of the inferential role in its account of content. Externalism denies internalism, and assumes that content is wide, or determined by the relation to the external world (Burge 1979; Putnam 1975), which may be naturally spelled out in terms of causal roles. Some externalists claim that representational contents is simply caused by representational targets (Dretske 1980; Stampe 1977). This view usually has problems explaining how mental representations can have contents and is exposed to the disjunction problem (Fodor 1992), which appears because in every misapplication of a concept, the concept is caused not only by its referent A, but also by its non-referent, say, B. But just because the content, *ex definitione*, is caused by the referent, it has to be assumed that the content is $A \vee B$. And with every error, the story repeats, thus the content is highly disjunctive (hence the name of the problem). Thus, it seems that simple causal accounts cannot specify the content, or satisfaction conditions of representations. Some externalists, however, appeal to the idea that misrepresentation is asymmetrically dependent on the correct representation, and thus avoid the disjunction problem (Fodor 1992). It remains unclear however how this idea of asymmetric dependence is supposed to explain mental representation at all (Bickhard 2008; Crane 2003).

The proponents of teleosemantics attempt to explain mental representation in terms of biological teleological function (Cao 2012; Dretske 1997; Millikan 1984). The notion of function is either etiological (Millikan 1984, 1989), or based on the self-maintenance of biological systems (Anderson and Rosenberg 2008; Bickhard 1993, 2008). Explaining mental representations in terms of teleological function allows teleosemanticists to account for misrepresentation, characterized as a representational dysfunction rather than as something that is not a representation at all (cf. Bielecka 2014; for an early but classical account, see Dretske 1986).

While teleosemantics is widely regarded as the most successful account of mental representation (Artiga 2016; Shea 2013), that does not mean that it has settled the issue. There is an ongoing debate among representationalists and anti-representationalists about the nature of mental representation and its role in cognition. There are two distinguishable kinds of representationalism: ontological and epistemic. Ontological representationalists claim that representations are real (Dretske 1986; Millikan 1984) as opposed to ontological anti-representationalists (Hutto and Myin 2013), while epistemic representationalists claim that they have a (merely) explanatory role (Dennett 1971; Sprevak 2013) as opposed to epistemic anti-representationalists (Chemero 2009).

Most ontological anti-representationalists are limited representationalists who reject only some kinds of representations: total anti-representationalism would be self-refuting if only it denied that there are any representations and was stated using one. Some anti-representationalists reject all mental representations except for propositional attitudes considered as statements in the language of thought (Fodor 1975). But Fodor's opponents reject the idea that such symbolic representations are really contentful (Cummins and Roth 2012; Stich 1983). Currently, Fodor claims that other kinds of representations such as mental images may also exist (Fodor 2008) but he denies that there is a role for contents understood in terms of specifying the features of targets, or Frege's Sinn (Fodor and Pylyshyn 2015). Interestingly, there is an affinity between this Fodorian view and anti-representationalists who reject mental representations in favor of basic indicators or detectors and external linguistic representations (Hutto and Myin 2013), or in favor of affordances (Garzon 2008) posited by James J. Gibson (1986) in his ecological psychology (but see a representational interpretation in: Bickhard and Richie 1983). Hutto and Myin claim that the problem of naturalizing satisfaction conditions cannot be solved (hence they call it a *Hard Problem of Content*), which is consistent with Fodor's diagnosis of the problem as well. Yet their opponents deny that detector representations — the only admissible mental representations, but devoid of content according to Hutto and Myin, or of meaning according to Fodor and Pylyshyn — should be understood as mental representations at all because they have no content (Cummins and Roth 2012; Gładziejewski 2015; Ramsey 2007), and argue for the representational character of structural representations. According to them, only representations based on resemblance relations can play representational roles in cognitive systems. But that also means that Ramsey and Cummins are anti-representationalist in the sense that they reject detector-like representations. Note, however, Ramsey's claim that structural representations are no longer present in contemporary cognitive science or cognitive neuroscience; however, Shagrir (2012) and Gładziejewski (2016) have found plausible evidence to the contrary.

2.2. Proposed research and novelty of the approach

The research proposed within the current project will be based on the account of mental representation already developed by Bielecka, which has turned out to be quite successful for criticizing attempts to characterize semantic phenomena (see, for example, the criticism of various solutions to the Symbol Grounding Problem, cf. Bielecka 2015, 2016). The account is based on teleosemantics and includes a model of system detectable-error. A satisfactory account of mental representation should fulfill desiderata listed below.

Metadesideratum I: The satisfactory account of cognitive representation is connected with empirical research.

I.1. The account elucidates the role of representation in describing, explaining, predicting, and controlling cognitive processes and behaviors.
I.2. The account is applicable to both linguistic and non-linguistic (e.g., perceptual) phenomena in human beingsandotheranimals.

I.3. The account is applicable to personal and sub-personal phenomena.

Metadesideratum II: The satisfactory account of cognitive representation is able to explicate the representational character of intentional notions.

II.1. The account is compatible with assumptions about the rationality of cognitive systems. II.2. The account is able to naturalistically explicate the notion of intentionality, that is:

II.2.1. It indicates the determinants of reference and content.

II.2.2. It shows why some representations can be bearers of truth.
II.3. The account offers an account of content (including the non-contingent relationship between contents and referent).

II.3.1. It does not exclude productivity of some representations. II.3.2. It makes misrepresentation possible. II.3.3. It makes system-detectable error, or misrepresentation for the system, possible.

The doctoral dissertation of the Principal Investigator, supported by the National Science Center under the program PRELUDIUM III, *The Role of the Concept of Misrepresentation in the Modern Theories of Mental Representations*, argued for a satisfactory philosophical account of mental representation in which intentionality can be naturalized, i.e., explicated in non-intentional terms. The Principal Investigator proposed the above list of desiderata that a satisfactory account of mental representation should fulfill. It was argued that if one rejected the possibility of misrepresentation (Perlman 2000), one would have to assume that any unsuccessful action is the mark of irrationality (as the agent could not achieve the goal of the action). This is why the notion of misrepresentation is necessary in order to avoid the consequence that all actions are based on accurate representations. This is what desideratum II.2 requires in detail: it does not, specifically, require that cognitive systems be rational in general. The last two desiderata (II.3.2 and II.3.3) are arguably the most difficult to fulfill. The dissertation examined the extant naturalistic philosophical accounts of mental representation, paying special attention to the possibility of misrepresentation being detected by the cognitive system itself. Because the most successful account did not turn out to be satisfactory, a sketch of a new account was proposed, namely a coherence-based account of system-detectable error, which is based on the assumptions of teleosemantic attempts to naturalize intentionality.

In the proposed model, mental representation is understood along the lines of teleosemantics (Millikan 1984), i.e., in a biological and evolutionary way. In the model, mental representations are cognitive states or processes of a cognitive system whose ability to represent is explained functionally: they have a representational biological function. To account for the functionality of representations, a hybrid account of biological function that is partly etiological (as in Millikan's (1984, 1989) theory of proper function) and partly autonomy-based (as in Bickhard's (2008) autonomy account of function), was adopted. The model also includes Millikan's model of a producer and consumer of representations. So, to give a simple example, a rat has a representational proper function to represent food if in the history of the organism a feature of olfactory apparatus has been selected for such a function. The bearer of a proper function is the producer system, or its olfactory apparatus, and the consumer, or its motor system. A misrepresentation is simply a dysfunction of a representing system.

The model assumes that the detectability of mental representations' adequacy is the necessary condition for a cognitive system to represent (for a similar argument, see also Bickhard 2008; Fodor 1992, p. 107). In other words, only systems which can misrepresent can represent. As a result*,* a coherence-based model of detecting representational errors by the cognitive system was proposed. This model is based on the idea of system detectable-error defended by Bickhard, extended by adding an inconsistency condition, which states that the adequacy of mental representation is available to the system as long as it can detect informational inconsistency. It was argued that in order to detect that a representation A is inconsistent with B, a cognitive system must be able to register an inconsistency of A, and the information or representation B must be available to the system. Furthermore, a cognitive system can to correct the misrepresentation X only if it can detect the misrepresentation X. To correct the error, it suffices for the system to have an ability to detect the difference in the degree of reliability of information A and B (for a similar account of error-correction, see Deacon 2012).

This idea has its predecessor in the Husserlian notion of harmoniousness of experience (Husserl 1989, p. 78). If two senses conflict with one another, for example, when Macbeth reaches for a hallucinated dagger, this conflict is resolved by assuming that the experience is no longer veridical. Obviously, the cognitive mechanism in place need not be conscious, and it requires that a given causal process has the ability to compare two individual physical tokens of a certain medium and check to ascertain if they are the same as each other.

The proposed model will be used to analyze anti-representational accounts of mental disorders. It should be noted that the model is different from the usual conceptual background of the debate in the contemporary philosophy of psychiatry, in which one of the important topics has been the nature of delusions: whether they should be considered as beliefs or not (Bortolotti 2010; Kapusta 2015; Schwitzgebel 2012). According to the phenomenological definition proposed by Karl Jaspers, delusions are false judgments held with strong certainty that are immune to contrary experience or argument and have impossible content (Jaspers 1963). His definition is similar to the one proposed under DSM-IV and DSM-5, according to which delusions are false beliefs based on incorrect inference about external reality that persists despite evidence to the contrary (American Psychiatric Association. DSM-5 Task Force. 2013). The notion of belief or judgment, while deeply entrenched in the philosophical tradition, and arguably present in folk discourse, is not necessarily the basic notion of mental representation, as the preceding discussion shows. The obvious problem with understanding delusions as beliefs is that they are immune to revision even when one is presented with contrary evidence or argument, and beliefs are usually understood as implying rationality (Dennett 1987). While this discussion has turned out to be philosophically very fruitful, the current project proceeds from a different starting point. There is no particular reason to take folk psychological categories for granted when studying the representational nature or aspects of mental disorders. There is also no need to require that all contentful representations of cognitive systems depend on their rationality, even if most cognitive systems may indeed be in some sense instrumentally rational. This is simply a different empirical issue that can (and in the case of psychopathology, even should) be studied to see whether a given cognitive agent displays bounded rationality (Simon 1956), Bayesian rationality (Oaksford and Chater 2007), or maybe even optimal rationality, or none, *etc*. Moreover, it is not necessary to assume that all representational states are propositional attitudes. To take a simple example: Love is a representational mental state, as it is directed towards someone or something. But it is not (necessarily) an attitude towards a proposition or state of affairs (Crane 2003, p. 25).

So instead of using the folk psychological categories, the project will appeal to the technical term *representation* as defined in the teleosemantic framework and studied in cognitive science or cognitive neuroscience (also as a sub-personal phenomenon). While it might turn out that propositional attitudes held by patients are always their mental representations, the current research does not need to assume this without further evidence. On the contrary, it may seem that in some mental disorders, patients confabulate unknowingly and constantly, and hence, they may produce linguistic utterances without really asserting them, which will mean that it could be difficult to ascribe them any attitude towards these utterances. In the case of severe retrograde amnesia caused by Korsakoff's syndrome, a patient may almost immediately forget an utterance spoken a couple of minutes earlier (for a striking classical case, see the description of Mr. Thompson in Sacks 1985). It should be noted that many anti-representationally oriented philosophers or psychiatrists presuppose that contentful mental representations are just propositional attitudes understood one way or another, and thus not sub-personal phenomena. But it could still qualify as a representation in a different sense. At the same time, there is no need to require verbal behavior in order to ascribe mental

representations in the teleosemantic framework, which makes it possible to study the nature of mental disorders of non-linguistic animals.

For this reason, the research team will extend the existing teleosemantic model of mental representation, and use it to study phenomena at various levels of the cognitive system's organization (or a distributed system that includes the patient, other people, and significant items in the environment). One need not presuppose that all kinds of hallucinations (for a review of various reports, see Blom 2010) have the same underlying cause. It should not be assumed that perceptual misrepresentations cannot be at the core of hallucinations just as well as perceptual beliefs can. It is an interesting and open question that asks what kinds of representations are involved—and how—in various kinds of mental disorders. The requirement that the cognitive system should be able to check the consistency of its representations, which is at the core of the model, can be used to distinguish misrepresentations, or representational errors (as errors detectable for the system) from non-representational errors. Because the model is open to further development, it should be extended during the investigation of psychopathology cases.

The research team will investigate some cases of psychopathology (or rather their theoretical accounts and descriptions of clinical cases) that have cognitive or emotional components. Special attention will be paid to their potential representational and non-representational features, while mental representing is understood in terms of the biological function of an organism that is able to detect its own errors via inconsistency. The answer to the problem will be given by describing and analyzing some paradigmatic cases of psychopathology. The analysis will focus on hallucinations, confabulations, cognitive impairments of empathy as pathological symptoms, as well as obsessive-compulsive disorders and psychoses as disease entities (in both humans and animals).

Hallucinations. *Hallucination* is sometimes defined broadly as a percept experienced by a waking individual, in the absence of an appropriate stimulus from the extracorporeal world (Blom 2010). But this definition would also include mental imagery, thus most authors require that hallucinations be involuntary (e.g., Beck and Rector 2003). Hallucinations can affect all sensory systems and there are a number of different kinds (for an accessible introduction, see Sacks 2012). They seem to be representational in the first place, as they have content which is available off-line. Offline representations are defined by their decouplability or detachment from their target (Clark 1997; B. C. Smith 1996). Furthermore, some of them can be recognized via an inconsistency in visual and tactile stimuli. An example of hallucination that can be explained this way is Charles-Bonnet syndrome. This is a complex visual hallucination in people with some impairment of vision and its content is bizarre in nature (they include figures in elaborate costumes, human beings of non-natural size, fantastic creatures, or extreme colours, which may partly overlap with real visual perception). Some Gibsonian anti-representationalists argue that hallucinations can be better explained by time-extended perceptual processes instead of decoupling (Garzon 2008). Hallucinations, as they argue, are passively experienced and there exists no external source of stimulation (Gibson 1970). Yet other anti-representationalists claim that hallucinations are like a real presence that occurs also in perception (Noë 2005).

While *prima facie* this kind of anti-representational account can work with kinds of hallucinations that the subject of hallucinations has seen before, it does not seem to be plausible for hallucinations of novel objects. Among them are also visual hallucinations, as in Charles-Bonnet syndrome, in which patients see bizarre figures that have never been seen before. To sum up, even if some kinds of hallucinations could be explained anti-representationally, many of them remain problematic for anti-representationalists.

Confabulations. Generally speaking, confabulation is a sort of pathological certainty about ill-grounded thoughts and utterances; confabulation involves absence of doubt about something one should doubt: memory, ability to move, ability to see (Hirstein 2005; Reuter 2014). In DSM-IV, confabulation is defined narrowly, as "the recitation of imaginary events to fill in gaps in memory" (American Psychiatric Association 2000, p. 443) and fits Korsakoff's syndrome well (in which a person's spatial, verbal and procedural memory is left intact but there are large deficits in declarative memory) but not other syndromes in which memory doesn't play such an important role. In DSM-5, confabulation is more widely defined, as a syndrome that appears in a variety of other syndromes, including those involving no memory problems, as anosognosia for hemiplegia (denial of paralysis), Capgras' syndrome (the illusion that an impostor has replaced a person close to the patient) and schizophrenia (American Psychiatric Association. DSM-5 Task Force. 2013).

Confabulations seem to be representational. There are two kinds of errors present in confabulations: 1. a patient gives a false response, so the ability to construct plausible responses is not retained; 2. a patient fails to check falsity of his response, so the ability to verify those responses is damaged. Hence, in some pathological cases, as in Korsakoff's syndrome or in dementia anti-representationalist is plausible. In such cases, the content of mental representation seems to be so vaporous that its role in the patient's cognitive life is questionable, especially in Korsakoff's syndrome when a patient feels free to produce words that he doesn't remember, it is natural to explain the syndrome with external (therefore, non-mental) representations. The anti-representational explanations in this sense could be narrative (Hutto and Myin 2013).

Cognitive impairments of empathy. The proper definition of *empathy* is still a matter of debate. Adam Smith defined *sympathy* (the predecessor of the contemporary notion of empathy) as the effect that is produced when we imagine that another person's circumstances are our own circumstances, and find their reaction to the circumstances to be reasonable (A. Smith 1761). A little consensus has emerged in the psychological literature about what counts as empathy, and shared emotions seem to be its essential component. Coplan and Goldie (2011) define *empathy* by appeal to its three central elements: affective matching, perspective-taking as other-oriented (not self-oriented), and self-other differentiation. It is assumed to be distinct from sympathy, compassion, or emotional contagion. According to the recent literature, empathy is no longer considered a unitary concept, but comprises at least two components: cognitive and affective (Decety and Meyer 2008; Dziobek 2008; Singer 2006). While the first concerns the ability to understand other people's feelings or thoughts, the second is responsible for the ability to emotionally 'resonate' with other people's feeling. Cognitive empathy is the ability to take someone's perspective or place oneself in someone's situation, that is to be able to understand and respond to another person's mental state.

Despite the discrepancy between the phenomenological analysis of a concept of empathy and the neuroscientific research on this phenomenon, two functionally-defined levels of empathy, lower- and higher-level, are considered in contemporary research. The lower-level is considered to be more affective in nature, the higher-level seems to be more cognitive. While some, like Amy Coplan, include both levels in their definition of empathy, others, like Murray (2011), consider the lower-level as a precursor of real higher-level empathy.

There is a body of research on the connection between psychopathy and the autistic and psychotic spectrum (Ashwin et al. 2006; Baron-Cohen 2011; Blair 2008; Corden 2008; Howard 2000; Humphreys 2013; Wallace 2008). Baron-Cohen emphasizes that people with autism lack cognitive empathy (preserving the affective component intact), while people with psychopathic personality disorder (borderline personality disorder, psychopathy, narcissism) lack affective empathy but have cognitive empathy intact, although some studies seem to contradict the latter claim (Harari et al. 2010; for a review, see Jeung and Herpertz 2014). It suggests that lack of affective empathy plays an important role in humans as moral and rational agents (but see (Bloom 2016)).

It is plausible that the cognitive impairment of empathy is representational, especially when the theory of mind hypothesis is assumed (Gopnik et al. 2001). So the ability to understand people's thoughts and feelings could be understood as mentalizing their cognitive states and feelings, which requires inferences (Singer 2006). While theory of mind hypothesis seems to fit cognitive empathy, it doesn't fit the affective one well, where the role of inferences seems to be limited. An anti-representational account might appeal to a sensorimotor simulation as playing a role in affective empathy. *Prima facie*, anti-representationalism seems to be more plausible if affective and cognitive empathy are not strongly interrelated.

Psychosis. According to DSM-5, psychosis is not a well-defined disorder (American Psychiatric Association. DSM-5 Task Force. 2013). Previously, it was connected with schizophrenia but not so strongly as in the previous DSM-IV (American Psychiatric Association 2000). Now catatonia is much more connected with psychosis. It can be characterized by such positive syndromes as delusions, hallucinations, disorganised speech and two negative syndromes: diminished emotional expression and avolition (Blanchard and Cohen 2006; Messinger et al. 2011). The contemporary attempts at classification tend to replace strict criteria (level, number, duration) with more loose ones, which are thought to more accurately capture the mechanism of psychosis (Heckers et al. 2013). Psychosis seems to be a very complicated disorder, and difficult to classify as either representational or not (but see Adams et al. 2013 for a representational and computational model). As for its delusional character, it seems more probable that it’s representational. The other symptoms however, such as emotional expression and avolition, are related more to emotions and their cognitive content is less obvious. Further investigation is required to ascertain whether emotional content of psychoses is really contentful.

Obsessive-compulsive disorder (OCD). According to DSM-5, obsessions are defined by (1) recurrent and persistent thoughts, urges, or impulses that are experienced as intrusive and unwanted, and that in most individuals cause marked anxiety and distress, (2) individual attempts to ignore or suppress such thoughts, urges or images, or to neutralize them with some other thoughts or action (e.g., by performing a compulsion). Compulsions are defined by (1) repetitive behaviors that the individual feels driven to perform in response to an obsession or according to rules that must be applied rigidly, (2) behaviors or mental acts aimed at preventing or reducing anxiety or distress, or preventing some dreaded event or situation; however, these

behaviors or mental acts are not connected realistically to what they are designed to neutralize or prevent, or they are clearly excessive. The obsessions or compulsions are time-consuming or cause clinically significant distress or impairment in social, occupational, or other important areas of functioning. They are not attributable to the physiological effects of a substance or another medical condition. The disturbance is not better explained by the symptoms of another mental disorder.

In a submitted paper, the members of the research team applied the proposed model of mental representation to mentally-ill non-human minds, in particular in obsessive-compulsive disorders (Braitman 2014). There are studies showing that some captive animals exhibit compulsive behavior (Dodman 2016). There is a famous case of a polar bear named Gus in Central Park Zoo, who would swim figure-of-eights for 80 percent of his waking hours (Grandin and Johnson 2010) to keep himself calm. Sometimes animal psychologists or ethologists would rather talk about abnormal repetitive behaviors than OCD (Holden and Travis 2010). This term, however, is misleading because it leaves out the possibility of a representational component. Traditionally, OCDs in non-human minds are understood in purely behaviorist terms, thus anti-representationally. However, the members of the research team argued that at least some OCDs might involve representational phenomena as well, which and under-diagnosed because of the sheer lack of verbal reports from animals. It is planned to investigate whether some aspects ofOCDs in human beings could also be understood in representational terms, or whether the anti-representational account is to be preferred.

Aphantasia. Aphantasia is a recently described disorder, which is not yet well classified or explained (Zeman et al. 2010, 2015). It is defined as a reduced or absent voluntary imagery. The terms used previously in related contexts include *defective revisualisation* (Botez et al. 1985) and *visual irreminiscence* (Nielsen 1962). Aphantasia could be representational because it is a dysfunction of visual imagining. In aphantasia, visual memory loss and deficits selectively disabling imagery are registered, that seem to have at least partly representational character. Interestingly, aphantasia does not yield easily to a Gibsonian explanation because there is no disability of perception at all. Some sceptics however claim that it is a mere fantasy case. Others claim that aphantasia is as doubtful as introspection simply because describing our inner lives is difficult and undoubtedly liable to error (Hurlburt and Schwitzgebel 2007). A more detailed preliminary study of this controversial phenomenon will be performed by the Principal Investigator before the beginning of the current project, as a co-investigator in Dr. Piotr Kozak's project funded by NCN under the decision 2015/19/D/HS1/02426.

After performing the case studies, the teleosemantic model of mental representation will be reassessed, and possibly enriched. Moreover, the anti-representational perspective on psychopathology will be analyzed in more general terms to see whether it could and should serve as a methodological starting point in psychopathology. While for human mental disorders, many assume that the representational approach of one kind or another is unavoidable (see, e.g., Poczobut 2014), in particular for disorders of self-representation (e.g., various kinds of anosognosia), the animal psychopathology assumes the opposite, and currently, most animal disorders are diagnosed merely behaviorally. It is therefore worthwhile to inquire whether parsimony considerations really license such an attitude, or whether an argument based on precaution principles could be made, to state that while the cost of false positives in the diagnosis of mental disorders is negligible, the false negatives in animal psychopathology are poised to cause unnecessary suffering. Moreover, one can argue that there are methodological reasons why the anti-representational attitude, even if supported by parsimony considerations, is not to be preferred: one needs to study the organism's relation to the environment, and that relation is best accounted for in semantic, and not just causal terms. This assumption lies at heart of contemporary cognitive neuroscience (Shagrir 2001), and it may play a role in integrating computational models in cognitive neuroscience with psychiatry. Note that recent computational models in neuroscience that appeal to predictive coding (Clark 2016) have made some initial attempts to explain psychopathology (Adams et al. 2013), and these models include a role for error-detection that seems in line with the proposed model of mental representation.

2.3 Impact of the project

Mental health problems are one of the main causes of the overall disease burden worldwide (Global Burden of Disease Study 2013 Collaborators 2015). According to the *Mental Health Atlas,* psychiatric disorders contribute to 13% of the global burden of disease (World Health Organization 2015). Mental health and behavioral problems (e.g., depression, anxiety and drug use) are reported to be the primary drivers of disability worldwide (Lozano et al. 2012). Major depression is considered the second leading cause of disability worldwide (Whiteford et al. 2013). Studies performed in the United Kingdom estimate that one in six people in the past week has experienced a common mental health problem (McManus et al. 2016). The global cost of mental illness is about 2.5 trillion dollars and is expected to increase to more than 6 trillion dollars (World Health Organization 2015). Those with mental illness are at high risk of developing other diseases, such as diabetes or cardiovascular disease (Kincaid and Sullivan 2014, p. 1).

For this reason, the project is not only important on theoretical grounds, it will provide a better understanding of the nature of different kinds of psychopathology, as well as a more precise understanding of the nature of representational phenomena that are revealed in psychopathology. Because clinicians primarily use common-sense terms to describe representational phenomena, this work may also contribute to the elucidation of the conceptual problems that occur in clinical practice. Psychiatry is a field without a unified conceptual or theoretical framework, and therefore any attempts to account for general regularities and integrate them with the results already obtained in cognitive neuroscience and cognitive science can contribute to its overall theoretical integration.

4. Research methodology

The work of the research team is guided by three philosophical approaches. These are: 1. a naturalistic approach to mind and cognition (Clark 2008; Dennett 1991; Millikan 1984; Thagard and Findlay 2012); 2. a historically-oriented, post-Kuhnian approach to philosophy of science (Friedman 2010; Hacking 2004; Wimsatt 2007); and 3. a naturalized approach in phenomenology (Gallagher and Zahavi 2008). Among various approaches in philosophy, phenomenology has proven especially sensitive to research on mental health. These three approaches converge in the proposed methodology. This means that it is not necessarily required just to analyze philosophical concepts, perform thought-experiments based on “intuitions”; rather, naturalistic philosophy summarizes and subsequently integrates the variety of existing work on specific topics to gain new insights into particular areas of research – in this case the systematic reassessment of anti-representational accounts of psychopathology, which fuses the best practices in cognitive science and teleosemantics. More precisely, the approach combines literature research, discussion with other researchers (scientists and philosophers alike), and the rigorous testing of new ideas. Whereas the Principal Investigator,

in her doctoral studies, focused on providing a systematic list of desiderata and a sketch of a teleosemantic account of misrepresentation (see e.g., Bielecka 2016); in this project, she will focus on the scientific literature and practice related to various kinds of psychopathology.

The research team also includes a clinical psychologist with a philosophical background and good knowledge of the history of psychiatry, which will be critical to the success of the project. Note that it has been repeatedly stressed that the use of statistical manuals such as the Diagnostic and Statistical Manual (DSM) to classify and understand mental disorders is highly controversial (Demazeux and Singy 2015; Kincaid and Sullivan 2014), and while it remains a useful starting point for the discussion, it cannot replace case histories and empirical studies. This is why the PI will consult the psychologist with the broad knowledge of various problems with classifying and understanding mental disorders.

At the same time, the approach will require rephrasing and paraphrasing the arguments presented by anti-representationalists with wildly different theoretical assumptions as there is no single theory of anti-representationalism (Kirchhoff 2011). This will require the use of the traditional method of paraphrasing, as used in the analytical philosophy.

This methodological approach has broader implications for the understanding of the role of philosophy in the academic arena. Philosophy should exclusively rely on reason, intuition, and reflection. In other words, philosophers ought not to adjudicate from outside the sciences and aim to correct scientists when they are confused or in error – quite the opposite. Philosophy should ‘get dirty’, productively engage with the sciences and ultimately seek integration of concepts, theories, and even methods on specific questions and problems. The relation envisaged here between philosophy and the sciences is, therefore, not one of conflict, theoretical separation and (occasional) correction but rather one of cooperation, deep engagement and (continuous) integration. At the same time, philosophy does not have to renounce its normative dimension; in particular, reflection on methodological and conceptual conundrums may lead to normative guidance for the subsequent research practice in the sciences.