ABSTRACT

Title of Dissertation: REPOSITIONING COGNITIVE KINDS

Aida Roige Mas, Doctor of Philosophy, 2022

Dissertation directed by: Distinguished University Professor Peter Carruthers, Department of Philosophy

This dissertation puts forward a series of theoretical proposals aimed to advance our understanding of cognitive kinds. The first chapter introduces the general debates that provide the philosophical underpinnings for the topics addressed in each of the following chapters. Chapter two compares and distinguishes between modules of the mind and mechanisms-as-causings, arguing that they should not be conflated in cognitive science. Additionally, it provides a novel "toolbox" model of accounts of mechanisms, and discusses what makes any such account adequate. Chapter three addresses the question of whether there is a role within the new mechanistic philosophy of science for representations. It advances a proposal on how to carve working entity types, so that they may include representational explanans. Chapter four offers an account of mental disorders, one that captures the regulative ideal behind psychiatry's inclusion of certain conditions as psychopathologies. Mental disorders are alterations in the production of some mental outputs (e.g. behaviors, beliefs, emotions, desires), such that their degree of reasons-responsiveness is extremely diminished with respect to what we would folk-psychologically expect it to be.

REPOSITIONING COGNITIVE KINDS

by

Aida Roige Mas

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Peter Carruthers, Chair
Professor Lindley Darden
Professor Eric Saidel
Professor Harjit Bhogal
Professor Bob Slevc

Dedication

For my parents and *àvia*

Acknowledgements

I am deeply grateful and indebted to many for their support throughout the years, not only throughout the writing of this dissertation, but throughout my life more generally. I am profoundly indebted to my parents, who always believed in me and encouraged me to pursue my path even if it was out of the ordinary. Also to my beloved *àvia*, to whom I dedicate this dissertation I promised to her on her last days. I am most indebted to Peter Carruthers, who is not only a brilliant philosopher and cognitive scientist from whom I learned a lot and who helped me tremendously in this project, but is the best supervisor I could have ever asked for. I don't think there is any mentor that is more kind, diligent, insightful and dedicated than Peter. I am very fortunate that he guided me so generously throughout this dissertation and my career, and I have accumulated a profound debt of gratitude to him. I owe many thanks as well to Lindley Darden, who has in effect come to be another advisor to me. Lindley has been incredibly supportive and encouraging of my work, revising and commenting on it countless times, and guiding me throughout with her knowledge of philosophy of science and biology. I am deeply indebted to both of them for my intellectual and professional development.

Many thanks are owed as well to Eric Saidel, whom I had many stimulating discussions with and has provided insightful comments on my work. I am also indebted to Harjit Bhogal and Bob Slevc, from whom I learned much and I am very grateful to count with as committee members. At the University of Maryland, College Park, I was fortunate to find an engaging and nurturing intellectual community. Many thanks to Sam Kerstein for introducing me to biomedical ethics and for his support

more generally. To Georges Rey, for many enjoyable conversations full of insights. Special thanks to Andrew Fyfe for his companionship throughout our shared doctoral journey. To him, Christopher Masciari, Julia Janczur, and Heather Adair, whom I am lucky to call friends. I am grateful as well to the many faculty members, colleagues and fellow graduate students from which I benefited during my graduate career, including Dan Moller, Elizabeth Schechter, Christopher Morris, Aiden Woodcock, Julius Schöenherr, Mike McCourt, Lia Curtis-Fine, Shen Pan, Kyley Ewing, Cody Gomez, Evan Westra, Jeremiah Tillman, Xintong Wang, Zhaoqi Hu, Ken Glazer, Louise Gilman, and many others. I also thank the DC History and Philosophy of Biology (DCHPB) group (including Makmiller Pedroso, Joan Straumanis, Kalewold Kalewold, and other past and current members) where I found a vibrant community. I am also deeply grateful to those who encouraged me during my academic journey, including faculty and colleagues at Universitat Autònoma de Barcelona (specially Daniel Quesada, Thomas Sturm, Josep Manuel Udina, and Èric Arnau Soler) and CSIC (Mario Toboso and Fanny Brotons). I also have countless people to thank for their comments as I have presented my ideas at conferences, workshops, reading groups and works-in-progress. A Fulbright-Spain fellowship supported me my first two years in the program and allowed me to take extra courses.

Last, but not least, I would like to give many heartfelt thanks to Andrés, for his patience, love and support throughout the making of this dissertation. Thanks as well to the rest of my family, especially Ferran, Sergi, padrí, and in loving memory of those who unfortunately, passed away while I was away.

Table of Contents

Dedication ii
Acknowledgements iii
Table of Contents v
Chapter 1: Introduction 1
1. How is the mind/brain composed? Two sorts of mechanisms in cognitive science 2
2. Can the New Mechanistic Philosophy of Science find a role for representations? 7
3. A material girl in a normative world: an account of mental disorders 11
Chapter 2. How is the mind/brain composed? Two sorts of mechanisms in cognitive science 17
1. Introduction 17
2. Modules or mechanisms-as-systems 21
3. An illustration of a mechanism-as-system: face recognition 33
4. Mechanisms-as- causings 36
5. An illustration of a mechanism-as-a-sort-of-causing: placebo analgesia 39
6. The relationship between mechanisms-as-systems and as-causings 43
7. How to treat work in progress 48
Chapter 3: Can the New Mechanistic Philosophy of Science find a role for representations? 52
1. Introduction 52
2. Representations in cognitive science 58
3. Fitting representations in the MDC account 62
4. Working-entity-hood 74
5. Getting it ARIght 90
6. Conclusion 95
Chapter 4: A material girl in a normative world: the extreme reasons-irresponsiveness account of mental disorders 96
1. Introduction 96
2. Two questions about mental disorders 98
3. What makes a disorder *mental*? 106
4. Considerations favoring my account of mental disorders 117
5. Why does psychiatry restrict itself to disorders of reasons-responsiveness? 123
6. Conclusion 126
Bibliography 128

Chapter 1: Introduction

One of the more useful distinctions in philosophy is that of tokens versus types: we distinguish concrete particulars from general sorts of things. Science is mostly in the business of providing explanations and predictions of general (as opposed to particular) phenomena, and it employs general taxonomical categories ("types" or "kinds") to do so. However, in cognitive science, what exactly can we say about those kinds? And what makes some such categories better than others for induction and explanation?

Instead of abstract, general recipes that don't capture the particularities of the cognitive domain, I believe that the work that remains to be done is closer to the scientific ground in which the discipline develops. Rather than providing a general account of kind-hood for the cognitive sciences, I believe we should answer this question on a case-by-case basis. Thus, my dissertation aims to provide three different —but related— contributions to debates involving cognitive kinds. The first paper distinguishes between two *sorts* of kinds that produce cognitive phenomena: modules of the mind ("systems" or "mechanisms") and mechanisms-as-causes. The second paper provides an account of how to carve out a mechanism's working entities such that these may include representational ones. The third paper provides a novel account of what sort of thing mental disorders are.

The aim of this introduction is to sketch the philosophical debates in which these three papers are situated. After providing the relevant philosophical background, I will briefly summarize the thesis of each paper as well as the contribution it makes to those debates.

1. How is the mind/brain composed? Two sorts of mechanisms in cognitive science

The question of mental architecture, or of what the underlying structure of the human mind/brain consists in, has involved at least two major debates.[1] The first debate concerns the type of processing by which the mind/brain converts inputs into outputs: classicists (e.g. Chomsky, Fodor, Pylyshyn) considered that this was done via symbol manipulation —analogous to symbolic computation in digital computers. Meanwhile, connectionists (e.g. Hinton, McClelland) held that this processing occurred via dynamic, parallel, and patterned activity in networks that connect simple processing devices; a proposal they saw as more biologically plausible (Dawson, 1998).[2] The debate around type of cognitive processing slowed as two things became clear: first, that classical architectures could be implemented by neural networks and vice-versa; and second, that the mind/brain could involve more than one type of information

[1] In these papers (and more generally), I assume that some version of physicalism is true.
[2] Arguably, a third contender appeared later on: embodied and extended cognitive science, which emphasized the acting and interacting in the world aspects of information processing.

processing, and that indeed it seemed to do so for different information-processing tasks.

This debate between classicists and connectionists paved the way for a second debate: one concerning the *functional* architecture of the mind/brain. It was, and still is, common for cognitive scientists to talk about the "modules," "systems," or "mechanisms" composing the mind/brain: long-term memory, face recognition, visual perception, and so on. This approach of breaking down the mind/brain into its functional components is informed by faculty psychology, and has as precedent Franz Gall's phrenology (1835). Both faculty psychology and Gall viewed the mind as compositional: i.e., as something that could be best understood when explained in terms of separate functions, powers, or faculties. In philosophy of cognitive science, during the 1990s and 2000s a major point of contention was the *extent* of, and the *characteristics* of, modules of the mind. Regarding the extent of mental modules, Jerry Fodor (1983) held that only the peripheral systems of the mind were modular: that is, only the input (perception and language) and output (action) systems were modular, while central cognition (higher-order processing) was not. Modular systems, according to him, are informationally encapsulated: they cannot rely on information held elsewhere in the mind during the course of its processing. In contrast, central cognition is isotropic: it had access to all domains and could potentially use any relevant available information. Fodor's view was countered by many authors who argued for massive modularity. Massive modularity entails that the entirety of the mind/brain is modular. For this to be possible, information encapsulation (as Fodor described it) had to go. Many authors disputed as well other aspects of Fodor's

characterization of modules: content domain specificity, strong localizability, shallow outputs, innateness, fast processing and automaticity (see e.g. Barrett & Kurzban, 2006; Carruthers, 2006; Coltheart, 1999).

Peter Carruthers provided, in his 2006 book, what I consider to be the correct account of modules —or at least, the closest to the notion of modularity actually used by cognitive scientists. He characterized modules as "isolable function-specific processing systems, all or almost all of which are domain specific (in the content sense), whose operations aren't subject to the will, which are associated with specific neural structures (albeit sometimes spatially dispersed ones), and whose internal operations may be inaccessible to the remainder of cognition" (Carruthers, 2006, p. 12). He distinguished between narrow-scope information encapsulation (Fodor's) and wide-scope information encapsulation —which is true of a system if it has access to some, but not all, exogenous information during the course of its operations. Carruthers (2006) also argued that comparative psychology alongside evolutionary considerations make the most plausible case for the massive modularity of mind hypothesis.

Independently of these developments, philosophers of science were identifying a plurality of types of explanation in the sciences. New Mechanists argued that explanations in certain special sciences were in fact *mechanistic*, a notion which they aimed to elucidate. At a minimum, mechanistic explanations work by modelling the mechanism in the world causally responsible for the phenomenon of interest.

Following Krickel (2019) and Nicholson (2012), I will classify their proposed accounts of mechanism into two sorts:

- *Mechanisms-as-systems*: mechanisms that consist of stable arrangements, structures, and interacting parts such that their combined operation produces predetermined outcomes. These will include early accounts of mechanisms by Stuart Glennan (1996, 2002, 2010), William Bechtel (2008a, 2008b), Bechtel and Robert C. Richardson (1993), and Bechtel and Adele Abrahamsen (2005).

- *Mechanisms-as-causes:* mechanisms that aren't object-like but process-like, in the sense that their operation is a manifestation of the causal processes involving several entities that act and interact. These will include the accounts of Peter Machamer et al. (2000), Phillis Illari and Jon Williamson (2012), and recently Glennan (2017).

Recent debates in the literature involving New Mechanisms have been about the suitability (or lack thereof) of one or other account to a certain (sub-)discipline within the special sciences. For example, does Machamer et al.'s (2000) account capture the mechanisms involved in explanations in evolutionary biology? (Skipper & Millstein, 2005). What about neuroscience? (Craver, 2007; Chemero & Silberstein, 2008). The assumption seems to be that if any such account is successful, it is so because it captures the specific type of explanation deployed in a given subdiscipline.

At this point, it is worth noting that debates about modularity and debates about mechanisms are related, even if I haven't seen them connected in the literature. Insofar as New Mechanists assume that the mind/brain is composed of mechanisms,

are they providing a genuine alternative to modules as functionally dissociable units, or are they presenting modules under a different guise? If the latter, which notion of modularity were they employing? Similarly, do discussions on modularity give support to one account of mechanism over the others? And, do all explanations of cognitive phenomena rely on modules, or on mechanisms?

Chapter 2, "How is the mind/brain decomposed? Two sorts of mechanisms in cognitive science" aims to bridge this gap, providing an answer to these questions. I argue that post-Fodorean modules of the mind (such as those characterized in Carruthers' account) can be identified with mechanisms-as-systems (such as those characterized in Bechtel's account), since both families of views place similar conditions on modules/mechanisms being functionally dissociable, stable architectural components of the mind/brain. However, not all explanations of cognitive phenomena involve identifying a module(s) performing its proprietary function: some cognitive phenomena are the causal product of the workings of different systems in ways that go beyond their proper functions. In other words, some cognitive phenomena are produced by mechanisms-as-causes, and as such, their production does not entail the existence of a stable, dedicated system for their production. Moreover, I argue that at early stages of research, scientists should use the "minimal notion" of mechanism, which allows one to cash partial explanations in mechanistic terms without carrying over metaphysical commitments on what the target mechanism is like.

This paper makes a contribution to at least two different debates: (1) the debates on modularity of mind, and (2) the debates on what is the "best" account of mechanisms for a given scientific discipline. I address (1) by highlighting the minimal conditions a system must have in order to be a "module," and pointing out how these overlap with the conditions proposed by the mechanists-as-systems hypothesis; and (2) by arguing that the proprietary domain of a discipline may involve the action of more than one sort of mechanism, and so it is misguided to favor a single mechanistic account as the blanket answer to explanations in that domain.

2. Can the New Mechanistic Philosophy of Science find a role for representations?

Central to philosophy of mind has been the mind-body problem: the question of how mind and body are (and can be) related and how they affect (and can affect) one another. Most contemporary authors favor physicalism —the thesis that only physical particulars exist, such that a physical duplicate of this world would be a duplicate *simpliciter*. Within physicalism, there is a debate about the ontological status of putative mental kinds: are they reducible to underlying physical ones, or on the contrary, do they constitute an autonomous domain?

The position that mental kinds are reducible to physical ones is known as "reductive physicalism." Two objections have traditionally been raised against reductive physicalism. First, the absence of bridge laws connecting the mental to the physical suggests that one domain is not reduceable to the other (Davidson, 1970). Second,

reductive physicalism entails that creatures with different brains cannot have the same mental kinds as we do. So, if mental states can be multiply realized, reductive physicalism is false (Putnam, 1967; Block & Fodor, 1972).

Within non-reductive physicalist theories, a popular position is psycho-functionalism. Psycho-functionalism can be construed as the conjunction of two theses: that relevant ontological kinds are those posited by the best cognitive scientific theories, and that functionalism is true. Since the first tenet is self-explanatory, I will elaborate on the second. Functionalism is the view that what makes a mental state, event, process or property[3] (for simplicity, I will just talk of "states" in what follows) the kind of mental state it is (e.g. pain) is its functional role —that is, its causal profile— in the system of which it is a part. An internal state of an individual is an instance of type of mental state if, given a certain input, it performs the relevant causal role in relation to other states of the nervous system, and is causally efficacious in contributing to the subsequent behavior of the organism that possesses it (Putnam, 1975). It is worth clarifying that mental states are not *abstracta,*[4] according to the functionalist. When one is characterizing things by their functional role, one is not describing nonmaterial entities —but merely omitting[5] certain implementational details from the

[3] Functionalism doesn't have a commitment as to whether states, processes, events, properties, etc. are the correct kinds or metaphysical units in the mind/brain. The idea is the same regardless of metaphysical unit: what makes a process/event/state an instance of a given kind is the role it plays in the system in which it is part.

[4] Here I am referring to the distinction between concrete vs abstract things. Concrete things are those that are located in space and time, or —for an alternative characterization— that have causal powers (e.g. an atom, this book). In contrast, abstract stuff does not have spatiotemporal locations, and couldn't possibly be physical. Others define abstracta by what lacks causal powers -i.e. can't cause anything (e.g. the number 7; English) (see Rey, 1997; Falguera et al., 2022).

[5] Abstraction or omission of details to characterize a kind is pervasive in the natural sciences, and if properly conducted is not problematic.

characterization. Items with the property of having a certain functional role (kind instances) *also* have physical properties[6]. Although mental states are *recognized,* in part, by the behavior to which they are a contributing cause, *they are not identical to* that behavior. Likewise, a state may be caused by, and play, certain causal role(s), but it is not *just* these causal role(s). Functionalism entails multiple realizability—the view that a single mental kind can, in principle, be implemented or realized by multiple distinct physical kinds. According to functionalism, there may be a one-to-many relations between mental and physical kinds.

Although psycho-functionalism does not, in principle, entail that the mind is representational (that is, that it contains intentional items that are about or refer to things), in practice its adoption involves accepting that the mind contains representations. This is because many successful theories in cognitive science involve explanations and generalizations ranging over representational kinds. Historically, some of the most successful theories in cognitive science have been developed from classical approaches, which emphasize the manipulation of physical symbols or representations according to some rules. Even those deriving from connectionist and embodied-extended approaches are representational (Dawson, 2013; Calvo & Gomila, 2008).

Yet, not everyone is onboard with granting the reality of functional and representational kinds. Some authors view theories involving representations in the

[6] Note that functionalism is compatible with interactionist dualism. However, most functionalists are physicalists (they think that there is nothing "over and above the physical", and that a physical duplicate of this world would be a duplicate simpliciter of this world), so I will just talk about physicalist functionalism in what follows.

cognitive science as suspicious, or even worse, deficient. For example, Gualtiero Piccinini and Carl Craver (2011) argued that functional (and other non-mechanistic) explanations in cognitive science are a temporary "patch," a product of our present ignorance of the underlying mechanisms operating in cognitive phenomena, which will inevitably dissipate as neuroscience becomes more and more integrated with psychology and other "higher-level" branches of cognitive science. Since Craver and Piccinini use functional kinds in their mechanistic models (e.g. "neurotransmitter", "selection pressure"), but not representational ones, it is to be supposed that what they feel uneasy about is that such explanations involve representations.

Carl Craver is one of the main articulators of the account of mechanisms that came to be known as "MDC", according to which:

Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions (Machamer et al., 2000, p. 1).

So, one might ask: does the MDC account require that the kinds involved in cognitive scientific explanations are neuronal, or otherwise physical?

A positive answer would situate MDC in the reductive physicalist side of the mind-body problem. To my knowledge, the only attempt to answer this question (Krickel, 2019) did not arrive to a convincing conclusion for either view. So, in this paper, I amend (or maybe interpret) MDC to be compatible with representationalism. I do so by developing an account on which properties determine whether an individual is an

instance of certain working entity-kind, which I call the "ARI" account —for Activity-Enabling, Robust and Individuating properties.

My paper situates the mechanistic account provided by MDC within the reductive vs non-reductive physicalism debate. Also importantly, it makes a contribution to the question of which properties determine working entity status. That is, what makes certain entities suitable to be part of a mechanism producing a phenomenon, and what excludes other entities from this categorization? In so doing, I answer Franklin-Hall's (2016) carving problem: the question of how to carve a mechanism's components in a way that is appropriate, non-arbitrary, and non-gerrymandered.

3. A material girl in a normative world: an account of mental disorders

Cognitive science is the discipline that systematically studies information-processing systems such as the human mind/brain. There is a related discipline that also studies the human mind/brain, but its goal is not merely to gain knowledge, but also to promote mental health: psychiatry. Psychiatry studies, classifies, and researches causes as well as possible venues for intervention in mental disorders. Yet, few taxonomical systems or kinds have had the assumption of realism questioned as vehemently as that of mental disorders has.

What sort of things are mental disorders? This question can be decomposed into three more nuanced ones. The first one asks about the metaphysical *nature* of mental disorders: are they natural kinds, socially constructed kinds, broken normal kinds,

pragmatic kinds, etc.? The second asks about what makes them *disorders*, as opposed to non-pathological features of persons. The third asks what makes them *mental* disorders, as opposed to, say, neuronal or somatic disorders. My paper addresses the third question.

What distinguishes the mental from the physical? In the philosophy of mind, there have been various attempts to find a feature or set of features that all mental states, processes, properties and so on have, and that all non-mental stuff lack. These feature(s) are sometimes called "the mark(s) of the mental" or criteria for mentality. The challenge is to provide such "mark(s)" in such a way that does not immediately presuppose the truth of a particular stance on the mind-body problem. There is still an ongoing debate about this, but two features have continued to be discussed since they were first proposed. First, phenomenal consciousness: it is the property which mental states have when it is like something to undergo them (Nagel, 1974). A problem for this proposal is that there are some mental states, such as standing beliefs or subliminal perceptions, that aren't conscious. Second, intentionality: the property of being *about* something, of being directed at something, of standing for something (Brentano, 1874). A difficulty for this proposal is that some qualitative states (e.g. tickles, pains) are mental states but they don't seem to be about anything (although representationalists have made a convincing case that those states also have content). Another difficulty is that it seems that some non-mental stuff can be intentional too: for instance, tree rings stand for a tree's age, and this introduction *is about* my dissertation. There are also other proposals, such as direct access, incorrigibility, and transparency, but I won't discuss them here.

The way psychiatrists have approached the mental, however, does not seem to be informed by those philosophical debates. Many conditions that cause distress and/or impairment affect phenomenal consciousness and intentionality, yet they are not mentioned in the Diagnostic and Statistical Manual of Mental Disorders (DSM) — most notably, those affecting perception such as prosopagnosia, blindsight or phantom limb syndrome, but also some affecting other cognitive functions, such as Alzheimer's, Lou Gehrig's disease, or migraines. The notion of the "mental" they use isn't characterized negatively either, that is, by capturing what does not have a known physical cause: Down syndrome is a mental disorder despite having a known physical cause (having a third copy of the 21st chromosome), and so is narcolepsy, which is caused by hypocretin deficiency.

The inclusion and exclusion of disorders in the DSM follows a complicated collective decision-making process, where social, political, technologic, economic and pragmatic considerations play important roles. It is plausible that the disorders currently in the DSM simply don't form a systematic, coherent hole, and that there are exceptions to any characterization of mental disorders —since there are no consistently applied criteria, metaphysic or epistemic, for something to be a mental disorder. Despite this, I think it is possible to provide a general characterization of mental disorders that captures the paradigmatic conditions listed in the DSM, drawing on a folk-psychological conception of the mind as containing reasons-responsive mechanisms. A mental disorder will thus be a significantly diminished (compared to a

non-disordered person) degree of the *reasons-responsiveness* of the mechanisms producing certain intentional states, actions and/or emotions.

Some clarifications: I don't take folk-psychology to be true (also, there isn't a single folk psychology but folk-psychologies, since some mental concepts are acquired through socialization and acculturation). I consider the "mental", so understood, as a generally useful fiction produced by the mind-reading system. If "the mental" vs "non-mental" distinction had some sort of interest-independent reality, we would probably fail to get the division right, given the large amount of evidence that we interpret, and to certain degree confabulate, the reasons behind our actions and those of others (Carruthers, 2011). I also hold the view that *all* mental phenomena are explainable causally. Moreover, this view of "the mental" entails that what falls upon its domain changes through time, to the extent that reasons-discourse and normative standards evolve alongside socio-cultural factors.

In philosophy, reasons-responsiveness has received the most attention from ethics. In a Frankfurtian compatibilist spirit, Fischer and Ravizza (1998) have an account of moral responsibility according to which a person is morally responsible for her behavior if she has guidance control over it, that is, if the mechanism producing these behaviors is *responsive to reasons*. Whatever the merits of this view to capture moral responsibility, I use this as the basis for an account of what makes certain disorders *mental* (as opposed to neurological or somatic conditions).

Many mental disorders can be characterized by such a lack of reasons-responsiveness: the severity of mood disorders, for instance, depends on the degree to

which one's mood is unresponsive to the things we consider provide reasons to change it. (Relatedly, an exclusionary factor for depression diagnoses is whether the person has an appropriate reason for experiencing its symptoms, such as the death of a spouse, in which case the persistent symptoms would not indicate a failure of reasons-responsiveness). A person with arachnophobia has an unfitting fear emotion to spiders that is resistant to sort of convincing that would modulate another person's fear. Delusions are resistant to evidence. This doesn't mean the mechanisms responsible for these states cannot be modulated: maybe they can, but venues for intervention would likely depend on causal, as opposed to reasons-giving, interventions (e.g. using antidepressants or cognitive behavioral therapy).

The connection between "the mental" in psychiatry and "moral responsibility" in ethics also explains why it is plausible that mental illness at least mitigates moral responsibility. The debate as to whether people with mental disorders are morally responsible for their actions is a contested one in ethics (see, e.g. Pickard, 2011; King & May, 2018; Kozuch & McKenna, 2016) and in law (e.g. Elliott, 1996; Kalis & Meynen, 2014). My account of mental disorders makes sense of why these debates exist.

Coming back to our original question: what can we say about cognitive kinds? My dissertation project tries to develop a piecemeal approach to what makes for a good kind in the cognitive sciences. The three papers comprising this dissertation aim, thus, to make contributions on this regard. In chapter 2, I distinguish between two sorts of mechanisms that should not be conflated in cognitive science. In addition, I provide a

"toolbox" model of mechanistic accounts of explanation: the adequacy of any particular account depends on both the target mechanism, as well as our current knowledge about it. In chapter 3, I vindicate the appropriateness of the New Mechanical Philosophy (and in particular, Machamer et al.'s 2000 account) to capture representational explanans, by providing a general account of how to carve working entities in the context of a mechanism. In chapter 4, I address a certain taxonomic kind —"mental disorders"— , with idiosyncratic characteristics that make them work differently than most other scientific kinds, and I provide an account of what guides the inclusion or exclusion of certain diagnostic categories in the psychiatric classification system. At the end of this dissertation, I hope, we will have reached a greater understanding of what cognitive mechanisms, representations and psychiatric disorders are.

Chapter 2. How is the mind/brain composed? Two sorts of mechanisms in cognitive science

1. Introduction

Cognitive science is a multidisciplinary scientific field; its domain of interest is cognition, broadly understood—how systems (especially human nervous systems) represent, store and process information. Cognitive science is not only interested in *what* our cognitive systems do, but also in *how* the relevant parts of the mind/brain make these things occur. A way to shed light on the workings of a cognitive system like the human mind/brain is by *decomposing* it to see *how* the relevant parts produce the phenomenon of interest. When the goal is to produce general explanations, this decomposition often involves *types* (as opposed to particulars).

In philosophy of cognitive science, two separate streams of literature that have taken on the project of answering the question "what should the mind/brain be decomposed into?" or "what are the relevant units when it comes to analyzing how cognitive phenomena are produced?". These two traditions have proposed different blanket answers: one tradition emphasizes the notion of "*modularity*". First introduced by Fodor (1983), this notion has been substantively revised in the hands of massive modularists (e.g. Cosmides & Tooby, 1992; Carruthers, 2006; Barrett, 2015).

Massive Modularists argued that the mind/brain is entirely composed by “modules” or "isolable function-specific processing systems" (Carruthers, 2006, p. 12). These systems are stable architectural components of the mind/brain, often characterized by appeal to domain specificity and a dedicated neural architecture (but without necessarily committing to the central component of Fodorian modularity, the encapsulation of modules). On the other hand, the New Mechanical Philosophy of science (after a widely-cited paper of Machamer et al., 2000, published some years after Glennan, 1996) proposes that many phenomena of the special sciences are the product of *mechanisms*. Some mechanists (e.g. Piccinini & Craver, 2011; Kaplan & Craver, 2011) have argued that all cognitive scientific phenomena are underlain by mechanisms-as- causings (or entities, activities and organization in causal continuity to produce the phenomenon). According to those mechanists, to explain cognition one ought to identify the relevant mechanism(s) involved.

So, is it modules or mechanisms that we should be decomposing the mind/brain into? The dispute is not merely a verbal one, they are genuine alternatives. Although one may be tempted to consider these accounts to capture two sides of the same coin, I argue that many times their targets are different in nature. What is described by each of those accounts is distinct and usually ought to not to be conflated with the other. At a minimum, there are metaphysical differences between the two: systems or modules are machine-like in that they *continue to exist* even when they aren't acting; while mechanisms-as- causings *only exist while they occur*.

This has implications for how we quantify and count cognitive components and for our treatment of deviant cases. For instance, consider John's ability to recognize faces. Treating face recognition as a module implies that John's face-recognition system is the same when he recognizes Mary and when he recognizes Thomas, and that same token module (m1) explains both these occurrences. On the other hand, from a mechanism-as-sort-of-causing standpoint, causes are only tokened while they occur, so each token instance of John's recognizing a friend (Mary, Thomas, Mary at a later time) is produced by a different *token* mechanism (call them mc1, mc2, mc3 …), which explains a particular occurrence of recognizing. Token systems (token modules) thus persist over time while token mechanisms-as- causings do not. Moreover, were John to acquire prosopagnosia, which is the inability to recognize faces as a result of a brain injury, the modularist would talk about a "damaged module" while the mechanist-as-a-sort-of-causing would talk about the different mechanisms that now underlie his response to facial stimuli[7].

This is not, however, the only difference to be found between mechanisms-as-systems and mechanisms-as- causes. The former are often universal among humans, with a distinct developmental path, a dedicated brain network, and a distinct proprietary function, while for the latter we may expect the involvement of distinct (and not so closely related) networks, greater variability in their starting conditions and/or parts involved, and invariance with respect to the production of the phenomena.

[7] The mechanisms will differ by belonging to different mechanism-types. The two different mechanism-types may be close together in similarity space, in the sense of having a good number of components in common, but they are still different because they produce different phenomena.

I argue that both approaches to mechanisms properly capture part of the functioning of mind/brain. I claim (section 2) that cognition involves modules or mechanisms-as-systems, which I illustrate by presenting the face recognition system in section 3. I argue that cognition also involves mechanisms-as- causings in section 4, which I illustrate by presenting a mechanism for placebo analgesia in section 5—a mechanism which, like many mechanisms-as-causings, doesn't have clear spatio-temporal boundaries nor dedicated processing structures. I address the relationship between both sorts of approaches in section 6. However, this raises questions about how to treat research in progress (section 7). Most research involves fragmentary sketchy models, that include some but not all the parts of the target mechanism. If both sorts of mechanisms (system-like and cause-like) exist in cognitive science, how do researchers decide into which category their proposed mechanism fits? How should they be treating their target system in the meantime?[8]

This leads us to the last section of the paper. I argue that we ought to understand hypotheses or proposals in the early stages as non-committal as to whether the phenomenon is underlain by a mechanism-as-system or mechanism-as-sort-of-causing. A good approach is to make use of Glennan and Illari's "minimal notion of mechanism" (2017) as a tool to start thinking about the target mechanism and incorporate the different findings. Their minimal notion has the advantage of being neutral with respect to the metaphysical commitments made by the other two accounts. I argue that at early hypothesis-generating stages, researchers are not saying

[8] This is important methodologically, as it would determine the methods researchers should choose to study the mechanisms of interest.

anything metaphysically "thick" about what is behind the phenomenon in question. It would be an error to assume, for instance, that every time evolutionary psychologists hypothesize that there is a "system" behind the observed phenomena of cheater detection, that what they have in mind a stable functional component or "system" in Carruthers' (2006) or Bechtel's (2008b) sense. Even if that ended up being the case, it wouldn't be epistemically justifiable to make that assumption until there is more evidence for the claim.

2. Modules or mechanisms-as-systems

Cognitive scientists often try to describe mental mechanisms, but they aren't always explicit about how they use myriad terms like "mechanism", "module", or "system". There seem to be at least two different ways in which they do so: very roughly, sometimes they treat them as systems (as in memory systems, language module, face recognition system, the visual system, etc.); other times, they mean the sort of causing that explains *how* something comes to be the case[9] (here, we could find the sequence of causally interacting parts that gives rise to instances of the decoy effect, McGurk effect, Thatcher effect or change blindness). Although the two senses are intuitively related (or may even overlap when a system is performing its proper function), they

[9] "How" in a causal sense, not in an evolutionary one.

capture different aspects of reality. In this section, I will discuss the first sense, i.e. the notion of mechanism-as-system or module.

When discussing face perception, Nancy Kanwisher wondered:

“Is face perception carried out by domain-specific mechanisms, that is, by modules specialized for processing faces in particular? Or are faces handled by domain-general mechanisms that can operate on nonface visual stimuli as well?” (Kanwisher, 2000, p. 759)

Like her, many cognitive scientists have discussed the units composing the mind/brain to be mechanisms in the “system” or “module” sense:

"An association may be found between tasks X and Y because the mechanisms on which they depend are adjacent in the brain rather than because they depend on the same underlying mechanism. Gerstmann’s syndrome is an example. It is defined by four very different symptoms: problems of finger identification; problems in calculation; impaired spelling; and left–right disorientation. It is improbable that the same mechanisms or modules are involved in all four tasks. What is much more likely is that these four symptoms depend on different mechanisms that happen to be anatomically adjacent in the brain.” (Eysenck & Keane, 2015, p. 20)

As these quotations illustrate, mechanisms are sometimes treated as *stable systems* composing the brain, with a specific function, which *perdure even when they aren’t acting*, and may be (perhaps loosely) *localized*.

This sense of modularity is implicit in Fodor's original formulation in his book *Modularity of Mind (*1983), and is the common ground between Fodor's view and those of other modularists who later disputed his characterization and provided their own (e.g. Carruthers, 2006; Coltheart, 1999; Barrett & Kurzban, 2006; Cosmides & Tooby, 1992; Pinker, 1997).

Fodor took modules to be systems characterized by being informationally encapsulated (that is, that during processing modules do not share, and cannot be affected by, information held anywhere else in the mind), as well as by exhibiting eight other characteristics[10]. According to him, modules are restricted to systems at the "periphery" of the mind —those that deal with perception, language and action. Many after him have argued that the entirety of the mind/brain is modular, a thesis known as "massive modularity". Most notably, Peter Carruthers (2006) argued that comparative psychology alongside evolutionary considerations make the most plausible case for the massive modularity of mind hypothesis. For massive modularity to be possible, modules cannot be informationally encapsulated —and unsurprisingly, this is the feature of Fodor's list that has received the most pushback. Other features

[10] Here is the full list of characteristics Fodor (1983, p. 37) proposed a module had to have "to some interesting extent" (meaning to an appreciable degree): (1) Domain specificity: modules are restricted to certain kind of inputs.(2) Mandatory operation: modules operate in a mandatory (or automatic) way. (3) Limited central accessibility: the operations and representations occurring within a module are not accessible (or accessible only in a very limited way) to higher cognitive processes. (4) Fast processing: modules generate outputs quickly.(5) Informational encapsulation: during its processing modules do not share, and cannot be affected by, information held anywhere else in the mind. (6) 'Shallow' outputs: the outputs generated are basic, simple. (7) Fixed neural architecture: modules are realized in a dedicated neuronal architecture. (8) Characteristic and specific breakdown patterns: modules can be selectively damaged, allowing for phenomena like double dissociations. (9) Characteristic ontogenetic pace and sequencing: modules are hard-wired in the brain and have a characteristic development. Among the features of this list, Fodor took information encapsulation to be a module's most important feature (Fodor, 2000), probably because he thought (mistakenly) it would solve the frame problem—the problem of how to restrict computations to what is relevant, "given that relevance is holistic, open-ended, and context-sensitive" (Shanahan, 2016).

of the list also got disputed: content domain specificity, strong localizability, shallow outputs, innateness, fast processing and automaticity (see e.g. Barrett & Kurzban, 2006; Carruthers, 2006; Coltheart, 1999). Nowadays, cognitive scientists largely operate outside Fodor's assumptions about the *extent* and *characteristics* of modules.

The important point for our purposes is that the family of modular views of the mind all share a common notion of module: modules are *stable systems* composing the brain, with a specific function, which *perdure even when they aren't acting*, and may be (perhaps loosely) *localized*. This basic characterization is a minimum common denominator among modular accounts, to which different authors add further characteristics[11]. This is also the common denominator among a certain family of accounts of mechanisms, as I will discuss next.

Daniel Nicholson' describes a family of theories of mechanisms that he calls "machine mechanisms": "systems conceived in mechanical terms; that is, as stable assemblies of interacting parts arranged in such a way that their combined operation results in predetermined outcomes" (Nicholson, 2012, p. 153). Beate Krickel (2019) labels it the "Complex Systems Approach" to mechanisms: accounts that "speak of mechanisms in terms of stable arrangements, structures, or objects" and "highlight the machine analogy in arguing that mechanisms are like machines" (Krickel, 2019, p.

[11] For example, Carruthers (2006) considers the following characteristics help us identify modules: modules have a function, something they are supposed to do with the information they receive. They are domain-specific. Modules are usually associated with particular areas of the brain, although they may be using a set of neural pathways that may be scattered in multiple brain regions. The processing of information that occurs in modules is largely independent of information stored elsewhere. Modules are entity-like in that they can be selectively damaged, allowing for double dissociations. Modules operate automatically, not at will. Most importantly, the mind/brain *contains* several modules; those are treated as its architecture.

22). These will include early accounts of mechanisms by Stuart Glennan (1996, 2002, 2010), William Bechtel (2008a, 2008b), Bechtel and Robert C. Richardson (1993), and Bechtel and Adele Abrahamsen (2005).

As I will spell out below, there are some commonalities between Complex Systems Approaches to mechanisms and accounts of modules, not only in their minimum common denominators discussed above, but also among other putative characteristics in maximally articulated accounts (e.g. Carruthers 2006 among modular accounts, Glennan 2002 and Bechtel 2008b among Complex System Approaches). The common features in these approaches make up what I will call “the mechanisms-as-systems” view.

The targets for both, Complex Systems and modular accounts, are stable entity-like systems that persist. For instance, Glennan talks about mechanisms as stable arrangements of parts that have dispositions: “Perhaps the most notable difference between the complex-systems and Salmon/Railton approach is that Salmon/Railton mechanisms are *sequences of interconnected events* while complex-systems mechanisms are *things* (or objects)" (Glennan 2002, S345, emphasis in original). Bechtel and Abrahamsen say that "[a] mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization" (Bechtel & Abrahamsen 2005, p. 423). I will start first by discussing the two features that come most to mind when discussing systems as object-like things: their spatial

localization, on the one hand, and their "function" or "predetermined outcomes"[12], on the other.

Mechanism-as-systems or modules are often associated with a particular part of the brain. Modularists typically speak of this as a module's "dedicated neural implementation", but that dedication isn't always exclusive. We should interpret identifications of a module with certain brain areas as an heuristic, not something to be taken literally. Bechtel describes it in the following way: "A common first step in such research is to attribute to a part of the system that produces a given phenomenon full responsibility for that phenomenon (in Bechtel & Richardson, 1993, we referred to this as direct localization). In the case of vision, this involved treating an area of the brain as the visual center. Such attributions of complex activities to single components of a system seldom turn out to be correct. As heuristic identity claims, however, they often contribute to productive research by facilitating the discovery that in fact the area is associated with only one or a few of the component operations required to produce the phenomenon." (Bechtel, 2008b, p. 89). In other words: spatial localization is a rough heuristic that helps to identify part of the implementation of a module —by looking at what areas selectively activate when performing a certain function. But more often than not, a module's functioning will require an integrated network of brain areas working together, sometimes including (sub-)modules scattered across the brain. For instance, face recognition comprises a brain network not only involving left and right fusiform face areas (which seem to be involved in

[12] The former is the term Bechtel and Carruthers use; the latter, Nicholson's.

discrimination of faces from non-social stimuli; Lopatina et al., 2018), but also occipital face areas (more involved in detailed recognition; Atkinson & Adolphs, 2011), the medial temporal lobe (specially the perirhinal cortex for familiarity-based recognition; Eichenbaum et al., 2007), and probably others. For another example, the affective system involves brain circuits that include nearly the entire brain: positive valence is processed in a network that involves subcortical regions of the ventral striatum linked with regions of orbitofrontal and ventromedial prefrontal cortex, while negative valence is processed in the amygdala and subcortical regions of the ventral striatum linked with the anterior insula, and cortically within the anterior insula and anterior cingulate (Berridge & Kringelbach, 2013; Yarkoni et al., 2011; Grabenhorst & Rolls, 2011). The point illustrated here is that the spatial localization heuristic is an over-idealization of a module's implementation: except for very simple and low-level sensory and motor systems, we can expect modules to comprise integrated networks of what may be different brain areas with specific sub-functions, which comprise a stable perduring system.

Modules or mechanisms-as-systems also have a proper function. This proper function is to be distinguished from things a token "merely does", does "by accident", or does when "malfunctioning". For instance, a clock is a mechanism-as-system for time-tracking, and an air conditioner, for cooling air. Even if they both take up space in one's apartment, that (taking space) is not their (proprietary) function, just something they do. The notion of proprietary function is not that of "mere causal role" employed by those who articulate mechanisms-as- causings: it is historical, and has an at least minimal normative dimension. There is something the module is supposed to do with

the information it receives, there is a set of phenomena mechanisms of that type are supposed to produce. While I do not want to tie modularity claims to a specific account of proper function, I will point out that the account must be one of the teleological family that can be traced back to Larry Wright. In Wright's characterization, the function of X is Z means (a) X is there because it does Z, and (b) Z is a consequence (or result) of X's being there (Wright 1976, p. 81). For Ruth Millikan (1984, 1989), Z is the proper function of X iff Z has had certain effects on how the tokens of X were "copied" or reproduced. These effects on the ancestors of a given token of X "have helped account for the survival, by continued reproduction, of the item's lineage" (Millikan, 2002, p. 8). For instance, hearts have been selected for by Darwinian natural selection (given that they did pump blood in "historically normal conditions"), and reproduced genetically. These are "proper biofunctions". In addition, the proper function of some modules may be a "secondary adaptation" or exaptation, if their preservation in recent evolutionary past is due to their doing G and not their F-ing they were originally selected for (Griffiths, 1993).

However, not all modules are adaptations, and thus not all functions of modules are grounded in natural selection: some modules acquire their function via learning mechanisms —for instance, those of the exact number system and the print reading system. Recently, Justin Garson (2017, 2019) has provided a historical account of function (which he calls the General Selected Effects, or GSE) that captures these two ways in which a mechanism comes to have a proper function. According to GSE, the proper function(s) of a mechanism is the activity that historically "contributed to its bearer's differential reproduction, or differential retention, within a population"

(Garson, 2017, p. 523). The "or" here is disjunctive: if a selection process has taken place resulting in differential reproduction or retention of a function, that is enough to make it a proper function —it doesn't matter if that selective process was natural selection, learning or competitive retention. For instance, suppose a learning process results in a neuronal disposition to Z to be retained, while another neuronal disposition to Y is eliminated, given a competitive process that takes place between them —a "zero sum game" (Garson, 2017, p. 532). In that case, Z-ing is the proper function of that mechanism. Something like this seems to be the process by which the print reading module is acquired: the left fusiform gyrus specializes in reading over the course of learning, resulting in the area known as "visual word form area" (VWFA) to acquire the proper function of recognizing letters and printed words, at the expense of conducting other tasks such as face processing. Supporting this hypothesis, Dehaene et al. (2010) and He et al. (2009) found that left fusiform responds to faces in illiterate adults, but such sensitivity is reduced when they learn to read. Among child beginner readers, Centanni et al. (2018) found that the greater the size of letter-sensitive cortex in left fusiform, the smaller the left face fusiform area (left FFA); and the greater the sensitivity of left fusiform to letters, the better the kid's reading ability.

Nicholas Shea (2018) has a similar account of proper function (which he calls "task function"). Roughly, an output F is a proper function of S if F is (i) a robust outcome function of S; and (ii) a stabilized function of S. An output F is a robust outcome function of S iff S produces F in response to a range of different inputs, and in a range of relevant external conditions (Shea, 2018, p. 55). An output F is a stabilized

function of S iff producing F has been systematically stabilized in at least one of the following manners: "(i) by contributing directly to the evolutionary success of systems S producing F; or (ii) by contributing through learning to S's disposition to produce F; or (iii) where S is an organism, by contributing directly to the persistence of S" (Shea, 2018, p. 64). The idea here is similar to Garson's: proper functions are grounded in processes that historically select Ss because they do F.

Like Garson and Shea, I contend that the processes of differential learning or selective retention are proper function-endowing. I suspect a majority of modules of the mind have acquired proper functions in this way: they start as partially innately-specified learning systems that become elaborated and built through domain-specific learning, a process that endows them with specialized proper functions.

The important point for our purposes is that the notion of "function" employed in mechanisms-as-systems is a teleological one, and not that of "mere causal role" employed by those who defend mechanisms-as-causings. This is true both for modularist authors as well as those who come from the mechanisms-as-systems traditions. For instance, William Bechtel and Oron Shaghir argue that, in order to describe a system computationally (in Marr's 1982 sense), one needs to specify what a system's *function* is, i.e. what it is meant to do, as well as its proprietary domain (see Bechtel & Shaghir, 2015).

Cognitive systems have proprietary domains; they are only sensitive to information of a certain sort. The "sort" of information a module is sensitive to is set by the properties of the inputs that reliably start a module's performance of its function, and not whether things in the world belong to a "sort". For instance, the proprietary domain of the face-recognition system are roughly face-like things, regardless of whether, in the world, these naturally conform to a kind. The generality or specificity of a proprietary domain doesn't have anything to do with how many instances of the thing there are, either. It is not the case that the proprietary domain of the face recognition system keeps expanding as more and more people populate the world.

It is important to emphasize that a module's sensitivity to its proprietary domain of information is due to its physical properties[13], and exempts the module from being actively working all the time. As an illustration, consider a photovoltaic cell. This cell is sensitive only to photons touching its surface, because of its physical-chemical properties and those of photons. Similarly, a module's physical properties may make it sensitive only to representations with properties of a certain sort.

That mechanisms-as-systems are *stable* and *entity-like* is also relevant in that it becomes possible to talk about a system as having a distinct development path, or being selectively damaged (thus allowing for double dissociations, i.e. cases when you have two related functions, X and Y, and after brain damage one subject can perform function X but not Y, and another subject can perform Y but not X; see

[13] "Physical" as opposed to "mental" properties, not the properties of physics. See Shaghir and Bechtel (2017) and Bechtel and Shaghir (2015) for discussion of the role physical or contextual constraints play in perception and computational tractability.

Shallice 1988). Carruthers (2006) took findings of double dissociation and/or a distinct development path to provide (non-conclusive) evidence for the existence of the module. We can get to know more about the proprietary domain of the module in question also by manipulating the input and then check whether the specific function is produced (that is, by manipulation; Woodward, 2003). Among the New Mechanists, Bechtel also discusses lesion research and double dissociations as providing *prima facie* evidence that the brain area in question is part of a certain mechanism-as-system. If damaging a certain brain area results in impaired functioning, then that would seem to suggest that the brain area in question was involved in the mechanism's operations (Bechtel, 2008b, p. 43). One thing to note is that lesion research can only provide evidence that something is a component of a system if we presume that the identity of the system is preserved despite the damage; that is, if we treat the system as by default perduring in time despite the fact that, in cases of damage, it may be missing some components and/or its functioning may be impaired.

The persistence of modules also makes them subject to selective pressures. An argument that is sometimes provided for the modularity of the mind is that modules are the best candidates to be units of selection for psychological traits. The idea here is that complex systems (like biological ones) need to be organized in a pervasively modular way, that is, as a hierarchical assembly of separately modifiable, functionally autonomous components, for the system to be constructed *incrementally*. Since the human mind/brain is a complex biological system that has evolved incrementally from animal mind/brains, it is plausible that it is modular (Carruthers, 2006).

In summary, mechanisms-as-systems or modules are *stable systems* composing the brain, with specific functions, which *perdure even when they aren't acting*, and may be loosely *localized*. Properties such as having a distinct developmental path, giving rise to distinct phenomena, having proprietary parts and operations, being distinctively sensitive to certain domain of information, and being dissociable add to the evidence for the mechanism in question being a system.

While there are some discussions as to whether modules are peripheral (meaning, only in perceptual and motor systems) or everywhere in the mind (including for higher cognitive functions), what seems clear is that cognitive scientists have provided numerous examples of mechanisms-as-systems: for example, memory systems, the language module, the visual system, and face recognition. In the next section, I will discuss the latter as an example of a mechanism-as-system.

3. An illustration of a mechanism-as-system: face recognition

When we talk about face recognition, we are referring to the identification of someone previously met using his/her face as the input. We are not talking about the recognition of a face as a face, as compared to another kind of object, nor of the recognition of facial expressions nor lip speech. The face recognition system is a mechanism-as-system which functions to recognize faces. Its proprietary inputs are face-like stimuli, and its output is a sometimes described as a sense of familiarity or recognition when we see a face of someone we know.

The most widely accepted model of face processing, by Bruce and Young (1986) separates the distinct processes that take place after encountering a face-like stimulus. Briefly: first, different codes are extracted from the input: pictorial codes (imagistic) and structural codes (more abstract). Those structural codes are then analyzed in order to obtain visually derived semantic codes (gender, age, personality), and identity-specific semantic codes (meaningfulness of a face). If the face is represented as familiar, the Face Recognition Unit (FRU) specific to that face becomes activated, producing a *feeling of familiarity*. This sense of familiarity takes place prior to the retrieval of any information about that person. Face-identity recognition ends, thus, with the activation of the FRU (or the lack of it) when presented with the stimulus face. Posteriorly, the FRU serves as a gateway to the information about that person, via stimulation of the Person Identity Node (PIN). In turn, that can activate the related Semantic Information Unit (containing information about the face's owner), which in turn can stimulate the Name Unit (containing that person's name). Starting from the early visual analysis and until the activation of a PIN, face-identity is processed and recognized independently of expression and facial-speech/lip-reading analysis.

A variety of evidence jointly supports the claim that face recognition comprises a mechanism-as-system or module. For example, there are *developmental impairments of face recognition as well as (non-pure) double dissociations*. For double dissociations, we have some cases in which brain damage has caused severely impaired facial recognition while recognition of other objects is around control levels (acquired prosopagnosia), while other patients exhibit severely impaired object recognition and normal levels of face recognition (object agnosia). There are also

documented cases of developmental prosopagnosia (see Bate, 2017, for an overview), which appears to be 'congenital' or 'hereditary'. Selective impairments of the face recognition system are better explained by assuming the functional dissociability of this system from others.

There is also evidence that face recognition has an *ontogenic developmental path of its own*. Babies exhibit a visual preference for faces (Valenza et al., 1996), which can be explained by both an innate facial module, and to what are good stimuli for immature visual systems (in terms of contrast, spatial frequency, top-heavy layout, etc.;). Face recognition develops and improves until one reaches approximately 10 years of age, becoming perceptually narrower (e.g. the other-race effect appears during this development). This effect is explained by both the face-specific module hypothesis and the expertise hypothesis (Hole & Bourne 2001).

In addition, there are face processing-specific *computational simulations*. Several computational systems have successfully emulated face-identity processing while exhibiting the same errors patterns humans do. Most are based on neural networks (e.g., Lawrence et al., 1997).

Finally, facial recognition *involves the activation of brain regions different from* non-facial recognition[14]. Neuroimaging studies have related it mainly to the fusiform face area (FFA) and the occipital face areas, which are regions that don't seem to have

[14] Maybe with the exception of the areas for visual expertise. There is some debate as to whether the FFA is specific to faces (but gets recruited for visual expertise in object recognition), to visual expertise (but gets recruited for face recognition), or both. See e.g. Tarr and Gauthier (2000), Gauthier et al. (2003), Kanwisher (2000).

other unrelated functions. But the module is likely to comprise a broader network of brain regions, including the hippocampus and the perirhinal cortex in the medial temporal lobe, some patches of the anterior inferior temporal cortex, and the amydgala (Lopatina et al., 2018). All this evidence supports the claim that face recognition is a mechanism-as-a-system or module.

4. Mechanisms-as- causings

So far, I have defended the position that mind/brain contains some mechanisms-as-systems. However, there isn't an entity-like system behind every cognitive phenomenon[15]. Sometimes, researchers individuate mechanisms of interest on the basis of *how* a phenomenon is causally produced—regardless of whether it forms a "system" or not. In this second sense, a cognitive mechanism is a specific "sort of causing":

"Researchers […] specify a mechanism *for the causal relationship* and combine the results from a variety of research questions." (Morling, 2017, p. 266, my emphasis)

Or:

[15] Doing so would be committing a version of what Bedford (1997) calls the "Not-The-Liver Fallacy," i.e. assuming that there must be a system responsible for every dissociable function of the organism.

"[T]here are many questions of mechanism, such as *How is object recognition accomplished by the visual system?*, that the lesion method, alone, is ill-suited to address." (Polsner, 2015, p. 56)

This is the second, mechanisms-as -causings notion. Explanations invoking mechanisms in this second sense address questions of "how does X occur?" by appealing to repeatable causal chains producing X. Here, mechanisms are individuated solely on the basis *what happens to produce a certain phenomenon*—they don't need to have natural boundaries or form a stable system[16].

This second notion is captured by Machamer, Darden and Craver's account of mechanisms-as- causings: "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al. 2000, p. 1; henceforth, "MDC"). In the rest of this section, this is the account I will be discussing because it is the most articulated, although it is not the only account of mechanisms-as- causings. The accounts of Illari and Williamson (1992) and recently Glennan (2018) also capture mechanisms in this sense, as what acts to produce a phenomenon. Krickel calls these "Acting Entities Approaches": they "assume that mechanisms are not objects but process-like, in the sense that they consist of actual manifestations of causal activities of various entities that interact" (Krickel, 2019, p. 25). Nicholson calls this the "causal mechanism" sense: "A step-

[16] According to Glennan (2007), there can even be one-off mechanisms.

by-step explanation of the mode of operation of a causal process that gives rise to a phenomenon of interest" (Nicholson, 2012, p. 153).

According to MDC, mechanisms have productive continuity leading to a phenomenon. Mechanisms themselves are composed of activities, entities and their spatio-temporal organization within the context of a mechanism-as-causing. Entities (i.e. the things that engage in activities[17] which have a robustness apart from their place in a mechanism) and activities (i.e. producers of change, specific "causings") are components of mechanisms (i.e. what produces or underlies a phenomenon), and they cannot be reduced to one another. MDC is explicitly and irremediably dualistic: "both entities and activities constitute mechanisms. There are no activities without entities, and entities do not do anything without activities." (MDC, 2000, p. 8).

We saw that modules or mechanisms-as-systems perdure, even when they are not acting. Craver and Tabery make very clear that mechanisms-as-causings do not do so: *"[e]ntities* (or *objects*) are not mechanisms. Mechanisms do things. If an object is not doing anything (i.e., if there is no phenomenon), then it is not a mechanism" (Craver & Tabery, 2019). That a mechanism is only such when it's producing a phenomenon doesn't mean that components of a mechanism cannot persist even when they aren't acting. Consider the mechanism of the action potential in neurons discussed in MDC 2000. The potassium channels, the axon, all of these continue to be

[17] It has been said that MDC individuates entities on the basis of their physical properties; that entities are concrete physical objects that exist in space and time (Krickel, 2019, p. 117). One might thus think that such account might have trouble accommodating explanations couched in intentional terms, but in fact it might turn out that dropping the requirement that entities are individuated on the basis of physical properties is a minor extension of the theory.

there, having “a kind of robustness and reality apart from their place within that mechanism” (Glennan, 1996, p. 53).

In summary: mechanistic explanation tells us how things work, by identifying entities, activities and their spatio-temporal organization, from initial or set-up conditions to finish or termination conditions. Mechanisms-as- causings *exist insofar as they are acting to produce phenomena*.

While these accounts were originally developed for other scientific disciplines, what seems clear is that cognitive scientists have provided numerous examples of mechanisms-as- causings: for example, the sequence of causally interacting parts that gives rise to instances of the decoy effect, McGurk effect, Thatcher effect, change blindness, or the well-known placebo effect. In the next section, I will discuss the latter as an example of a mechanism-as-a-sort-of-causing.

5. An illustration of a mechanism-as-a-sort-of-causing: placebo analgesia

The “placebo effect” is an illustration of the sort of case that captures nicely a mechanism-as-a-sort-of-causing. Placebos are sham treatments that are provided to patients, who believe them to be clinically effective. Placebos are well known to improve health outcomes: they account for as much as 50% of the effectiveness of analgesics, and they are also known to improve the immune response, motor outcomes in the case of Parkinson disease, depression, and so on (Barrett et al.,

2006). The phenomenon of interest is the improvement in the patient's condition following the administration of placebo.

Placebo effects are so robust that when a new medical intervention (for a condition for which there isn't a standard of care) is tested in a two armed, randomized clinical trial, a placebo is always given to the control group. Importantly, the improvement patients experience with the placebo cannot be attributed to the medically inert substance or intervention they receive, nor, in double blind studies, to expectations of the clinical staff. Instead, it results from the patient's anticipation that the intervention will help (Wager & Atlas, 2015). In other words, the improvement in patients' conditions when given placebos is *not caused by the physical or chemical properties* of the sham substance. Ingesting such a substance in another context, for instance, one in which the patient knows s/he is taking a sugar pill, has no effect. The placebo effect is caused by a mind/brain mechanism that has as a component the belief or expectation that the substance received will help.

There are a multiplicity of mechanisms that may come under the label of "placebo effects". Among these, the mechanisms behind placebo analgesia are among the most researched (Kong et al., 2007). In this section I will introduce some of the several mechanisms posited to underlie such pain relief.

Let's start by taking a step back and looking at normal pain processing in people without prior expectations. The brain contains several pain-responsive regions, some of which process the sensory components of pain (including the posterior insula, SS1 and SS2), others its badness (e,g, the anterior insula, thalamus and the anterior

cingulate cortex). When signals from across the body transporting information about noxious stimuli arrive, these areas are aroused, and the pain-responsive regions contribute to the generation of feelings of pain and the derived motor responses. Some even talk about a "neurologic pain signature" using fMRI on the basis of patterns of activation of those brain regions (see Wager et al., 2013; Reddan & Wager, 2018)

Placebos work by causing a reduction of the activity in these regions. First, the person needs to be given the placebo intervention in a way that mimics the treatment ritual (Sanders et al., 2020). The contextual information in the ritual —whether it is via verbal suggestions, retrieval of previous therapeutic experiences in similar settings or with similar interventions, observations of others getting treated, etc.; see Colloca (2019)— generates expectations of pain reduction. When someone has been primed to expect an improvement of their health state, their brain patterns are different from those of someone who is in a neutral state. Believing and/or expecting to receive an improvement in health leads to an increase in the activation certain areas of the prefrontal cortex and nucleus accumbens, which are believed to be responsible for appraisals of context and its meaning. Their activation leads to several changes. First, a partial suppression of the pain signals coming from the medulla carrying information about the type and intensity of the pain. The amplitude of event-related potentials produced in response to the painful stimuli is reduced, resulting in a decrease in experienced pain level (Wager & Atlas, 2015).

In addition, the activation of those areas in the prefrontal cortex results in the production and release of endorphins, a type of endogenous peptides that bind chiefly to opiate receptors in the brain, which results in a decrease in experienced pain. That some modulatory effects are mediated by endogenous opioids is well-stablished since a classic study (Levine et al., 1978) blocked placebo anesthesia effects by administering patients naxolone, an opioid receptor antagonist. Other neurochemicals mediating modulatory effects are dopamine, endocannabinoids, cholecystokinin (CCK) or serotonin, which also open possibilities for intervention (Frisaldi et al., 2020).

There are good reasons to believe that the analgesic effect of a placebo is a mechanism in the sort-of-causing sense. The placebo effect (and its sibling, the nocebo effect) seem to be a by-product of the more general evaluative learning system, which modulates expectations of goodness or badness based on previous cues. The formation of placebo responses has been accounted for as a general form of reward learning (de la Fuente-Fernández, 2009). It seems that the evaluative learning system is an adaptation, which at least in its associative (as opposed to cognitive) form we share with other animals. There is evidence of placebo effects mediated by Pavlovian conditioning occurring in e.g. rats (Herrstein, 1962). As I discussed in section 2, particular applications of a module's function do not qualify as proper functions themselves, unless they have been selected for in another way —e.g. via learning processes. But the learning involved in placebo is not selection of a function (i.e. it is not preceded by selective retention of a specific function to be performed in contexts related to pain and medical settings), but rather just another learning of cues

and their likely values — something to be expected given that module's proper function.

Consistent with this, the interventions that can be made on the effect size of placebos have to do with modification of the cues or their linked rewards, not with placebos qua such. For instance, both the effect size and the duration of a placebo response are often equivalent to the active treatment being studied (Tuttle et al., 2015). This seems to conform to a Woodwardian intervention, in that changing the magnitude of a component of the mechanism produces effects on the magnitude of the phenomenon (Woodward, 1997). As mentioned, it is also possible to block analgesic effects by administering the relevant receptor antagonists, which also have been found to reduce reward-learning based on e.g. food (Galaj & Ranaldi, 2021).

6. The relationship between mechanisms-as-systems and as-causings

Thus far, I argued that the mind-brain involves at least both mechanisms-as-systems or "modules", and mechanisms-as- causings. A complete cognitive science would probably involve modelling of both "sorts"[18]. Despite the superficial similarities, the two notions involve different metaphysical commitments. Modules have a continued

[18] I don't want to exclude in principle that there may be more "sorts" of mechanisms besides these two. Glennan (2017), for instance, allows for "one-off" mechanisms. Although I contemplate the possibility of one-off mechanisms, cognitive science is mostly interested in general explanations for general phenomena, so I won't discuss unique cases here.

existence, even when they aren't actively working. On the other hand, mechanisms-as-causings are temporally tied to the production a given phenomenon. Confusing the two is making a category mistake—it is confusing a continuant with a causal occurrence.

There are additional differences between modules and mechanisms-as- causings. I illustrated some with the examples discussed in 3 and 5, but let me elaborate a bit further on some other differences we can expect.

One such difference will be whether it's possible to change the range of circumstances in which the mechanism can operate. Modularists like Carruthers define a module's *domain* by the range of inputs suitable to turn the system on, not by module-independent input kinds[19] (2006). For example, the proprietary domain of the face recognition system are face-like objects that are processed by the face recognition system. A sometimes-overlooked fact is that token modules can change their proprietary domains through time. For example, there is evidence that during a child's development, the phoneme recognition system becomes increasingly specialized, resulting in increased sensitivity to phonemes of one's language, at the expense of increased insensitivity to those of other languages that cross-cut the phonetic space differently. For another example, the exact number system slowly expands its proprietary domain during development, as children learn to use number

[19] In contrast, for Fodor "domain specificity has to do with the range of questions for which a device provides answers (the range of inputs for which it computes analyses)" (1983, p. 103). While these notions are conceptually different, extensionally they point I make above remains.

words to represent number concepts (Libertus et al. 2016, p. 208). The domain of some modules can also change during adulthood: the face recognition system, for example, can be trained to also process Greebles (Gauthier & Tarr, 1997). In contrast, it is impossible that the range of circumstances in which a particular mechanisms-as-causing can operate changes. Once the explanandum phenomenon is fixed, the world fixes what are the background and set-up or starting conditions of the mechanism-as-causing (Craver & Tabery, 2019).

Relatedly, modules usually have *a dedicated brain network*, whether that is scattered across the brain or localized primarily in one area. In contrast, cognitive mechanisms-as-causings may have a lot more variation in their constituents. To mention a few: a mechanism-as-causing may be composed of a mixture of brain networks that aren't regularly that causally interconnected (as in the mechanism for spontaneously producing a joke); or by just a very small part of the brain in a way that doesn't follow natural boundaries (as in the mechanism of action potential); or by things outside the skull (as in the mechanism by which Otto came to know where Moma was, which includes his notebook, to use Clark and Chalmers' 1998 example). Part of the reason for this large variability in what realizes a mechanism-as-causing is that the selection of phenomena to be explained for the mechanist is, to a certain extent, interest-dependent.

Let me articulate what I mean by that. "Phenomena", for MDC and the theorist of mechanisms-as- causings, is an epistemic category, not a metaphysical one. Anything can be a phenomenon of interest; it is the scientists' choice to delineate it—including

making a decision about its *range of application* and *degree of detail*. While the researcher has a certain freedom in choosing the phenomenon, ultimately it is objective facts (about what produces that phenomenon) that establish what the underlying mechanism(s) for it is (are). Once a phenomenon is chosen, the explanatory work is to find the mechanism in the world causally responsible for it. In the best of cases, there is a single mechanism-type producing the phenomenon-type that the researcher is interested in. In other cases, however, lumping or splitting may be necessary. In any case, the mechanism is stablished by what it does. In contrast, the modularist and the theorist of mechanism-as-systems speak of a system's "proprietary function" (see discussion in section 2). The notion of "proprietary function" applies to a more restricted set of cases than that of "phenomena". A system's *function* is the task it is meant to perform; not everything a system *does* when it is operating qualifies. In contrast, a mechanism is delineated by what it does.

Despite these differences between modules and mechanisms-as-causings, there is a sense in which they overlap. Suppose that we fix the phenomenon in such a way that it overlaps with a module's proprietary function. As I mentioned before, once the phenomenon is fixed, the mechanism responsible for it is also delineated. Then, in that case, when a token module is performing its proprietary function (let us designate that function with the letter X), then that module is a token mechanism-as -causing for phenomena X, *while it is doing X.* That is, during the time in which the module is actively functioning as it is meant to, that system is a mechanism-as- causing for phenomenon X. In that case, it may be unnecessary to distinguish which of the two notions of mechanism we are using.

Nonetheless, it is important to keep mechanisms-as-systems and mechanisms-as- -causings distinct. Doing so allows for more clarity on quantification of mechanisms, as well as how to treat broken ones.

Regarding quantification, mechanisms-as-systems and mechanisms-as -causings have different ways to count tokens of a kind. If we are counting mechanisms-as- causings, then we would look at the phenomena being produced. A mechanism occurs only so when it's producing a phenomenon. Every time a mechanism is causally responsible for a phenomenon, then that is a token mechanism. When it stops acting, then we no longer have a mechanism. On the other hand, mechanisms-as-systems are objects that persist through time regardless of whether they are in operation or not. Here, sameness of mechanism seems to be determined by sameness of parts (although the exact conditions for identity are not clear; think of ship of Theseus-style problems). In any case, mechanisms-as -causings greatly outnumber mechanisms-as-systems.

Lastly, I want to highlight that the notion of "broken normal" or "damaged mechanism" only makes sense under modular approaches that take mechanisms to be systems. A module has a proprietary function, whether or not it carries it out. But mechanisms-as-causings exist only so insofar they are producing phenomena. If they are not producing anything, they don't.

In a nutshell: the set of modules and the set of mechanisms-as-causings involved in cognition are not co-extensive; while modules can be regarded as mechanisms-as-causings while they are performing its proprietary function, there are mechanisms-as-

causings (sometimes recurrent mechanism-types, like the placebo one) that aren't modules.

7. How to treat work in progress

In sections 2-6, I reviewed how mechanisms-as-systems and mechanisms-as-causings are different. However, it is worth taking a step back to focus not so much on complete explanations —cognitive science has few of these anyway— but on research in progress. Most research involves fragmentary models, in which we know some but not all the parts of the target mechanism. If both sorts of mechanisms (system-like and cause-like) exist in cognitive science, how do researchers figure out into which category their target mechanism fits? How should they be treating their target mechanism in the meantime?

Eric Hochstein has recently argued that working scientists should be "adopting deliberate metaphysical positions when studying mechanisms that go beyond what is empirically justified regarding the nature of the phenomenon being studied, the conditions of its occurrence, and its boundaries." (Hochstein, 2019, p. 579). In his opinion, it is acceptable for scientists to make metaphysical commitments at the onset of experimental investigation, provided that we are willing to revise these commitments on light of the evidence: "our metaphysical commitments might eventually need to be revised after a great deal of empirical work has taken place" (Hochstein, 2019, 588). However, in my opinion, making a metaphysical determination whether the target mechanism is a system or a causing before the

evidence is conclusive would only make scientists talk past each other. This can be seen, I believe, in discussions regarding the boundaries of the mind/brain[20] as well as mechanisms hypothesized by evolutionary psychologists[21]. The metaphysical commitments of our theories must correspond to what is warranted by the evidence, not to a philosophical demand to keep our metaphysics tidy.

Instead, I think here (at early research stages) is where we should employ the minimal notion of mechanism, formulated by Glennan and Illari:

"A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon" (Glennan & Illari, 2017, p. 2).

What makes this notion *minimal*? To start, it is meant to capture the main dimensions of the mechanistic framework in general, without tying it to more specific accounts. The different sorts of accounts can be seen as *varieties* of the minimal notion of mechanism, which add to it further details (and further commitments) of the account in question. Importantly, the minimal notion is not metaphysically committed to mechanisms being systems or causings, or constitutive or etiological, or persistent or one-off.

Back to my proposal. We should understand hypotheses or proposals in the early stages—when scientists speak of "mechanisms", "systems", "modules", "causal

[20] Adams and Aizawa (2001, 2012) and Rupert (2009) make similar points.
[21] See, for example, Atran (2001).

factors", "contributing factors", etc. behind a phenomenon—as non-committal as yet about whether the phenomenon is underlain by a mechanism-as-system or mechanism-as-sort-of-causing. At early, hypothesis-generating stages, researchers are not saying anything metaphysically "thick" about what is behind the phenomenon in question. It would be an error to assume, for instance, that a range of behaviors that would have been adaptive are always produced by a dedicated, innately-channeled "system" in Carruthers' (2006) sense. Even if that ended up being the case, it wouldn't be epistemically justifiable to hold that assumption until there is more evidence for the claim. But, they are definitely produced by (some) mechanism(s) or another.

The New Mechanistic philosophy provides us with a toolbox to aid discovery. Which account to employ depends on our epistemic state regarding the target mechanism, as well as its nature —whether it is a persisting functional component or a sequence of causally interacting parts. At beginning research stages, the minimal notion is the one that should be employed; every other account of mechanisms (and modules) is thicker. Each makes concrete proposals as to which properties would count as evidence that, in reality, there is a mechanism-as-system or a mechanism-as-a-sort-of-causing producing the phenomenon. For example, for MDC, it would be changes in a phenomenon when you intervene on a part; for modules, it would be developmental and pathological evidence. Each account offers heuristics to help guide discovery, precisely because they are *thicker*.

If I am right, cognitive science *must* make use of more than one sort of account of mechanism, because those accounts capture the variety of mechanisms and of stages of knowledge about them. None of the accounts, in isolation, is adequate to the different stages of epistemic progress in discovering a mechanism, nor can they do justice to the diversity of mechanisms we find when it comes to cognition. Part of the literature has treated mechanistic accounts as each having its own proprietary scientific field. I am proposing that within the same field (for instance, cognitive science), there may be more than one sort of mechanism. If more than one account of mechanism applies to cognitive science, maybe the same is true for other scientific fields as well.

This in turn puts an interesting twist as to how to understand the plurality of accounts of mechanisms available. If I am right, there is no winner-takes-all, even for a particular scientific discipline. An account may be appropriate for some mechanisms but not others; this is why we should view the New Mechanistic Philosophy as a toolbox of accounts. The criteria for choosing a given account of mechanism should not be merely a matter of scientific field, but a matter of the nature of the particular mechanism out there in the world one is investigating.

Chapter 3: Can the New Mechanistic Philosophy of Science find a role for representations?

1. Introduction

Within the philosophy of science there have been competing views on what explanation is. For example, the classic deductive-nomological model considered that to explain a phenomenon we must describe the initial conditions C, and the law-like generalizations L, from which to deduce the phenomena to be explained. However, it was soon evident that this universalist proposal posed some requirements on explanation that may be unachievable in the special sciences[22], where exceptionless generalizations are rare and individual variation is to be expected. Instead, explanations in those sciences take a different form: scientists often appeal to mechanisms behind phenomena, and see their work as that of discovering these mechanisms. For instance, while discussing research goals in a popular cognitive psychology textbook, Daniel Levitin writes:

[22] The "special sciences" is an umbrella term first introduced by (Fodor, 1974) to include "higher level" disciplines such as psychology and the social sciences (economics, sociology...). These days, the term is generally thought to include as well cognitive science, life sciences such as biology, and in general any scientific discipline that doesn't seem to have exceptionless laws —in contrast to what is thought to occur in the "hard sciences" such as physics.

"*Explaining* behavior [..] requires more than just a knowledge of causes; it requires a detailed understanding of the mechanisms by which the causal factors perform their functions" (Levitin 2002, p. 116, his emphasis).

He is not alone in considering that explanation in certain disciplines involves discovering a phenomenon's underlying mechanisms. The use of the notion of "mechanisms" in cognitive science is pervasive[23], and it was only a matter of time before philosophers caught on to that fact. In the year 2000, Machamer, Darden and Craver (MDC henceforth) provided a philosophical account of scientific explanations involving mechanisms, originally intended for molecular biology and neurobiology. They said:

"Mechanisms are entities and activities organized such that they are *productive* of regular changes from start or set-up to finish or termination conditions" (Machamer et al., 2000, p. 1, their emphasis).

This captured the traditional association between explanation and causation: to explain an event or phenomenon is to identify the mechanism causally responsible for it. Moreover, their insistence that "a mechanism is always a mechanism of a given phenomenon" (Darden & Craver, 2013, p. 52) respects the traditional distinction between explanandum (the phenomenon to be explained) and explanans (the mechanism behind it and its component parts) (Nicholson, 2012).

[23] For instance, a search for the key word "mechanism" in abstracts and titles of Wiley's journal "Cognitive Science" returns around 1200 papers published between 1990 and today; a similar query on Elsevier's "Cognitive Psychology", about 500.

Other authors have offered alternative notions of mechanism. To mention a few, Glennan (2002) presents mechanisms as complex systems that produce their behavior by interaction of their parts, while Bechtel and Abrahamsen (2005) claim that a mechanism is a structure performing a function by virtue of its components. As MDC's is one of the most well-articulated positions, their account is the one I will discuss for the rest of the paper, paying particular attention to how it relates to cognitive science.

The MDC account of mechanisms was originally developed for molecular biology and neurobiology, while leaving the door open for its applicability to other special sciences. Later, it was extended to neuroscience (Craver, 2007, p.vii) and cognitive neuroscience (Craver & Kaplan, 2011). In 2011, Carl Craver and Gualtiero Piccinini generalized the applicability of mechanistic explanation to the whole of cognitive science. In their paper, they seem to suggest that all *good* explanations in the field of cognitive science are mechanistic. They distinguish between successful explanatory models, which according to them are mechanistic (in the MDC sense); and functional (and other non-mechanistic) explanations in cognitive science (as those employed by Fodor, 1968; Dennett, 1978; Cummins, 1983). They regard the latter sort as a temporary "patch": a product of our present ignorance of the underlying mechanisms operating in cognitive phenomena, which will predictably dissipate as neuroscience becomes more and more integrated with psychology and other "higher-level" branches of cognitive science.

This is rather a surprising stance. On the one hand, MDC tell us to "follow the science" when it comes to finding out which activities there are. Other work by some of the MDC authors provide strategies to discover mechanisms and their activities (e.g. Darden & Craver, 2002; Darden, 2002, 2006; Darden et al., 2018), treating the discovery of activities as part of the scientific enterprise —to the point they have been criticized for not providing a philosophically informative account of activities (Psillos, 2004; Godfrey-Smith, 2009; Franklin-Hall, 2016). On the other hand, Craver and Piccinini's (2011) proposal entails excluding *computing*, *imagining*, *recognizing*, and other representation-involving activities from "bona fide" activities —even as they are used in cognitive scientific models!

In response, some authors have looked for counterexamples: cases of seemingly successful explanation in cognitive science that don't seem to be mechanistic in nature (e.g. Silberstein and Chemero, 2012, 2013; Meyer, 2020; even non-MDC mechanists have provided such examples: Glennan et al., 2022). In my opinion, the debate initiated by Piccinini and Craver's (2011) paper will inevitably come to a standstill: since unfortunately, there are few, if any, complete explanations when it comes to cognitive phenomena, it seems premature to take a stance on what they generally *ought to* look like.

More interesting, I think, is the question of whether MDC's analysis of explanation is compatible with current explanatory practices in cognitive science. *Representations* are common explanatory taxonomical categories in cognitive scientific models. This fact should not be surprising, given that despite numerous attempts,

representationalism —roughly, the view that many cognitive states and processes are most satisfactory modeled in representational terms—hasn't been successfully contested. Representationalism was historically disputed by some connectionists, first, and then by some dynamicists and sensorimotor theorists, but all these failed to provide an alternative way to capture important psychological and behavioral regularities –those that conform the "systematicity challenge"[24] (Fodor & Pylyshyn, 1988) and/or pertain to "representation-hungry" phenomena[25] (Clark & Toribio, 1994). The consensus seems to be that explanations sidestepping representations could not work outside explanations of individual occurrences, and maybe some special cases excluding "higher-order" cognition (Verdejo 2015). In addition, all these "alternative" views (connectionism, dynamic systems theory, and sensorimotor / embodied / extended approaches to cognition) are compatible with representationalism, thereby not offering a real substitute.

Given the abundance of representations in the explanans of cognitive phenomena, and Craver's (2006, 2007) insistence that they are mere "filler terms" or "metaphors"[26],

[24] The systematicity challenge is the challenge to explain the existence of systematic interrelations among cognitive capacities (e.g. the interrelations between the capacity to think "JOHN LOVES MARY" and the capacity to think "MARY LOVES JOHN") "without presupposing that the underlying processes are causally sensitive to the constituent structure of mental representations" (Fodor & McLaughin, 1990, p. 183).

[25] "Representation-hungry" phenomena are those that fall into one or both of the following categories: (1) phenomena which require sensitivity to objects or properties that are absent, counterfactual, or non-existent; and (2) phenomena which require "[selective sensitivity] to parameters whose ambient physical manifestations are complex and unruly" (Clark & Toribio, 1994, p. 419).

[26] Craver (2007) employs representation (entity) and *representing* (activity) in a few places of his book, but points out that terms such as "represent", "encode" or "process" are black boxes, question marks or filler terms that "are innocuous when they stand as place-holders for future work or when it is possible to replace the filler term with some stock-in-trade property, entity, activity, or mechanism (as is the case for ''coding'' in DNA), [but] are barriers to progress when they veil failures of understanding" (p. 113). When discussing spatial cognition, Craver and Darden (2001) include the following quote from Healy (1998, p. 133) with which they seem to agree (emphasis added): "The idea of a cognitive map, first proposed by Tolman (1948), has been an important and influential

one might wonder whether the MDC account of mechanisms itself is incompatible with representations. Can MDC's account capture the representational explanans employed by, e.g., psychology? Is MDC's account of mechanisms compatible with representationalism? Or, are we witnessing an explanatory approach that could revive the representationalism vs non-representationalism debate, after all these years?

I will argue that the MDC account of mechanisms can, in fact, incorporate representations as components of mechanisms —or, at the very least, that the MDC account can be easily amended to incorporate them, once the conditions for working-entity-hood are articulated.

One of the central points of my paper will thus be to provide an account of working-entity-hood for types —that is, an account of how to carve working-entity-types for general mechanistic explanations, which are the majority of scientific explanations (section 4). To do so, I will start in section 2 by briefly arguing that representations are central and indispensable when it comes to explaining cognitive phenomena. In section 3, I explore whether the MDC account can accommodate representation types that are individuated by their content as part of their ontology of mechanisms. In section 4, I provide a novel account of mechanistic entity-types, which I then discuss in section 5. Section 6 concludes the paper.

---

stimulus to research. *But it is really more a metaphor than a theory*. Research on path integration, landmark use, the sun compass and snapshot orientation … attempts to specify more concretely exactly what makes up a 'cognitive map' of space". The rest of Craver and Darden's paper discusses only non-representational entities and activities (maybe with the exception of *associating*). Craver (2006) treats *learning*, *encoding* and *retrieving* "memories" as part of the top behavioral-organismic mechanistic level, which is still a not well-understood "sketch".

2. Representations in cognitive science

Cognitive science is the interdisciplinary scientific field that studies information processing systems, such as the mind-brain. “Information” serves a demarcation role between cognitive and non-cognitive (say, merely physiological) phenomena and explanations. Trivially, everything carries information[27] about what it is causally related to; cognitive science, thus, uses a thicker notion of information; one that is often referred to by the term “representation”, which I will now unpack.

A representation is an information-carrying structure. Put differently, a representation is an object, event or state with semantic properties (reference, truth-conditions, content, truth-value, etc). Representations are intentional (a term introduced by Brentano, 1874): they are *about* something, are directed at something. For example, a representation of elephants *is about* elephants. Intentionality is a candidate to be a “mark of the mental”, what distinguishes representations and sets them apart. It is now common to distinguish between a representation’s physically realized state or structure, often referred to as the *vehicle* of representation (such as a state of the human brain), and the object or state of affairs represented, often referred to as the *content* of representation (such as what I am thinking about) (Dennett, 1991). The term “representation” captures this content-bearing vehicle, and refers to both its structural and its semantic properties.

[27] In the sense that “it is an indicator of”, or “reduces uncertainty”; see Shannon (1948).

Cognitive science assumes there are representations in the mind/brain because they best explain a range of behaviors. Consider, for instance, Edward Tolman et al.'s (1946) famous experiments on rats' maze learning. Rats got first trained to navigate a route (C) through a maze leading to food at (G). When placed in a novel maze with C blocked, rats disproportionately chose the path that would lead them straight to where the food should have been. If getting to the food was a mere matter of repetition, of simple association between stimulus and response as behaviorists argued, rats would be equally likely to choose either of the alternative routes offered. But they take the more direct route instead, because they *represent* where the food should be in their environment (Rey, 1997). Many decades of research suggest that this is done through the use of a cognitive map, a non-propositional (Camp, 2007) internal representation of the spatial environment that guides action, and is supported by the hippocampus and related structures (Epstein et al., 2017).

Another well-known example is the placebo effect: a beneficial health outcome produced by a medically inert intervention. Such a beneficial effect is not a product of the intervention's intrinsic features —since the placebo itself is designed to have no therapeutic value— but rather is caused by the expectations or beliefs surrounding an intervention. Such appraisals of the intervention as "apt to improve my condition" are commonly regarded as involving a representation, since the patient-subject displays sensitivity to a property whose physical manifestations are an open-ended, unruly disjunction: placebo effects occur with topical creams, pills, injections, surgeries, or even just possessing an analgesic. Moreover, the property of "being apt to improve my condition" is not even instantiated in placebos, although the patient-subject

believes it to be. Placebo effects seem thus to be "representation-hungry" phenomena (Clack & Toribio, 1994). According to Wager and Atlas (2015), the lessened degree of pain experienced occurs following an appraisal of the intervention as suitable to reduce one's pain, which modulates the brain's response in areas that usually process the affective aspect of painful stimuli.

These examples (and many more that could be found in the literature: stereotype bias, decoy effects, delay discounting, and so on) illustrate three things. First, representations are often[28] part of the *explanans* of general phenomena. A mouse takes a certain route on a maze because it represents the food as well as their current position on a cognitive map. That cognitive map legitimately explains the mouse's behavior. If I were to intervene to change the mouse's cognitive map (e.g. by moving the food and re-training the mouse, or by temporarily knocking out its place cells encoding the food's location), there would be a modification of its behavior. This relation between a mouse's cognitive maps and their behavior is causal and explanatory according to the interventionist approach favored by mechanists (e.g. Woodward's 2003).

Second, representations are part of the explanans both in *horizontal* and *vertical* explanations. Horizontal explanations are explanations of phenomena in terms of distinct states, events, processes and so on that temporally precede a phenomenon

[28] But not always; we may be interested, for example, on how a particular representation came to exist; in that case, we would be treating the representation as an explanandum.

(Bermudez, 2004) and cause it[29]. They are closely tied to the notion of an etiological mechanism, where a mechanism produces a phenomenon when it involves a causal sequence terminating in an end-product (Darden & Craver 2013; Craver & Tabery 2019). The placebo effect occurs when an initial propositional attitude towards a certain content is part of a causal chain of events producing a modification of how painful stimuli are affectively appraised; it is, thus, a horizontal explanation involving a representation. Vertical explanations, in contrast, can be broadly construed as explanations of what grounds a given phenomenon; they are explanations of a phenomenon in terms of its component features (Drayson, 2012; Bermúdez, 2004). Mechanists use the language of mechanisms underlying a phenomenon, or constitutive mechanisms, to refer to these (Kaiser, 2017). A mouse's capacity to navigate a maze can be vertically explained by an underlying mechanism that has a cognitive map as one of its components.

Third, these are *general* explanations: explanations ranging over a class of phenomena, not explanations of an individual occurrence. As such, they often involve *types*: general or repeatable sorts of things[30], as opposed to token or individual things. There is a repeatability to instances of the same kind (I will be using "kind" and "type" interchangeably in what follows). For instance, a cognitive map is a type of

[29] Although Bermudez doesn't make the causation claim, in what follows, I will treat mental causation (including causation by representations) as a possibility in principle, and I will bracket causal exclusion worries aside. Assuming physicalism, one may hold the view that causation really occurs at the level of the fundamental physical realizers of every token representation, and nothing I will say here aims to preclude this possibility. The reasons I am treating mental causation as unproblematic will become clearer in section 5.

[30] I am using "things" here broadly, in a way that doesn't commit me to a given metaphysical category. There may be types of states, entities, properties, activities, processes, and so on.

representation occurring in multiple mice. Representing something as apt to cure me is also a kind of way I (and others) can represent something to be. General explanations usually involve types both in their explanandum and explanans. These types do not need to be natural kinds in order to play such roles[31], provided they are apt to capture relevant non-accidental regularities.

As I discussed in this section, cognitive scientific explanations are often general explanations involving representation-types in their explanans. Since MDC's is an account of (also, but not only) general scientific explanations, it's natural to wonder how they fit together. So, where could one locate representations in MDC mechanisms?

3. Fitting representations in the MDC account

In the last section, I provided some examples of general phenomena whose explanation includes representations. These representation-types may be individuated on the basis of their structural properties (as in, visual representations), their semantic properties (as in, representations of something as apt to cure me) or both (as in, representations of one's location in a cognitive map).

[31] There are arguments that at least some such types are natural kinds, which I do not necessarily endorse. For instance, Smortchkova and Murez (2020) argue that certain representational kinds are natural; Michaelian (2011), memory systems; Davies (2016), color constancy; Izard (2017), basic emotions.

In this section, I will tackle the question of where (if at all) representations fit within the MDC ontology —which comprises mechanisms, entities, activities, and organization.

Are representations *mechanisms*? According to MDC (2000), mechanisms make things occur, they are productive[32] of phenomena. MDC describe mechanisms as working from start- or set-up conditions to termination conditions; other than when in operation in between those conditions, mechanisms don't persist —even if some of their components do. Metaphysically, an individual mechanism ceases to exist at the point where it is no longer operating; while a representation may persist (e.g. stored in a memory system) even if it is not participating causally in anything. Thus, representations aren't mechanisms, since representations are continuants and mechanisms are occurrents.

The same can be said about *activities*, which MDC define as the producers of change, sorts of causings. The MDC account of mechanisms is a causal-pluralist minimalist view (Godfrey-Smith, 2009) in that it treats activities as specifics sorts of causing, of which science must say what they are and what features they have. In contrast to representations, which are continuants, activities are occurrents that cease to exist when they are no longer producing any change.

Spatio-temporal *organization* seems important to determine the content of some representations, such as those of the cells in visual sensory area V1. V1 organization

[32] Whether that productivity is understood causally, as in the case of productive mechanisms, or constitutively, as in the case of underlying mechanisms.

replicates the spatial arrangement of the retina, and by extension, of the visual field. Each cell in V1 has a unique receptive field that only covers a relatively tiny area of the visual field, and they are positioned so that cells with neighboring receptive fields occupy adjacent locations along the cortical sheet (Patel et al., 2014). However, such organization alone doesn't give us content; entities (e.g. neurons) and activities (e.g. firing) need to be involved. Moreover, topographic and sequential ordering of neurons seems to be restricted to early sensory and motor areas. So it seems that spatio-temporal organization by itself is insufficient to fit representations.

Can *phenomena* be representations? That was the suggestion Beate Krickel seemed to make in the last chapter of her 2018 book. However, this sidesteps the question. Phenomena is a synonym for "explananda", and thus an epistemic category, not an ontological one: anything can be a phenomenon. What counts as a phenomenon depends on our interests and what we single out as in need of explanation. Certainly we may be interested on how representations are produced. However, representations are often appealed to in order to explain other phenomena (as in the navigation and placebo examples of the last section), and in those cases they play the *explanans* role, not the explananda one. So we are still in need of a place to fit explanans representations in the ontology of MDC mechanisms.

This leads us to the last and most plausible candidate: are representations *entities*? To avoid ambiguities, let's be clear that the question I am addressing is whether representation-types can be entity-types, as general explanations (the most common explanations in cognitive science) range over types. For MDC, "entity" is a technical

notion, different from the umbrella term used in some metaphysical debates to denote everything that exists. *Entity* is an ontological category in its own right, and distinct from *activity*, which makes MDC metaphysically dualistic (Krickel, 2019; Machamer, 2004). Machamer et al. (2000) define entities as the bearers of properties. Entities have relatively stable clusters of properties (Craver, 2007, p. 131), such as "crucial sizes, shapes, orientations, and locations" (Craver, 2001, p. 60). This way of characterizing entities is reminiscent of the way "substances" have traditionally been understood as the foundational elements of reality that "stand under", e.g., ways of being; in fact, when discussing the terminological choices of MDC, Peter Machamer says that they chose the term "entity" because it "seemed to carry fewer historical and philosophical presuppositions than "substance"" (Machamer, 2004, p. 27).

What are the properties entities have? Following the philosophical tradition, "properties" broadly denominate features, characteristics, predicables or determinate ways something may be; whether intrinsically or extrinsically (Orilia and Paoletti, 2020). Any object has an infinite number of properties, once we include extrinsic or relational ones; so do mechanistic entities. However, properties aren't all equally important. Presumably, a subset of an object's properties are relevant to its being a working part of a mechanism, while others are irrelevant.

The issue is what sorts of properties are relevant for working-entity-ness in the context of a mechanistic explanation; that is to say, which properties should we look at in order to carve the working entities in a mechanism. In the context that concerns us (of cognitive scientific explanations containing representations in the explanans),

the question boils down to this: do these properties have to be physical (in the sense of relating to the vehicle of representation *qua* physical structure), or could semantic properties (those relating to representations' content) suffice to pick out a mechanistic entity-type?

Unfortunately, Machamer, Darden and Craver do not give a clear answer to this question[33]. Like physicalists in the philosophy of mind, who consider that there is nothing "over and above" the physical, MDC argue that particular entities need to have a particular physical realization (in the mechanists' preferred terminology, they must be corporeal *objects*; see Craver et al., 2021). But (unless one is a nominalist) general explanations range over types or kinds of entities: for instance, the general mechanism of synaptic transmission includes working entities that are a certain way (e.g. entities of the kind *NA+ channel* acting to transport entities of the kind *ion* across the cell membrane). At a minimum, a kind (type) is a repeatable; something that can be exemplified in multiple individuals. For example, individual sodium (NA+) channels are instances of the kind *NA+ channel*.

What determines whether an individual is an instance of a given kind are its properties[34]. Properties are what we should be looking at for carving purposes. For

[33] For instance, they say "Traditionally one identifies and individuates entities in terms of their properties and spatiotemporal location" (MDC, 2000, p. 5), but it's not clear whether that is merely a historical fact or something to be expected given constraints on what can count as a working entity.

[34] In traditional natural kind essentialism, there is a (natural) property or properties that all and only members of the kind share, that makes them the kind of thing they are. In Boyd's Homeostatic Property Cluster view, some properties of a cluster that reliably co-occur are what determines the kind of thing it is. Non-natural kinds are also determined by their properties: for example, a bone is fractured if it has the property of being totally or partially broken; and being designed for the function of opening cans is what makes something a can opener.

instance, we should find out whether something is composed of H_20 molecules to know if it is water; whether it was designed to open bottles, to see if it is a bottle opener; and a protein's subunits and attachment to the cell's membrane to assess whether it is an instance of a sodium channel.

Given MDC's requirement that *individual* working entities are physical objects, the question is whether the MDC account of mechanisms requires that entity-types are carved by corporeal properties (e.g. shape, having a membrane) —or, if, on the contrary, non-corporeal properties (such as functional profile or semantic content) could do the job.

Let us call the first option [CKK]:

[CKK]: The properties by which working mechanistic entity-types should be carved are corporeal.

"Corporeal" here means tangible; it involves, at a minimum, having a determinate spatio-temporal location[35]. "Carving" is a term originating in Platonic metaphor, and here it is intended to be neutral on the issue of whether the taxonomical categories resulting from such "carving" are interest-dependent or practical categories constrained by the way the world actually is, or whether they are natural kinds corresponding to the world's "joints" or "seams".

[35] A better definition may be: properties that we cannot conceive being instantiated without having a determinate spatiotemporal location.

At first glance, [CKK] seems a plausible interpretation of the MDC account of mechanisms, given some mechanists' writings. Craver describes entities as spatially bounded objects (Craver & Wilson, 2006, p. 88). Kaiser and Krickel assert that, on the MDC view, entities "are conceived as material objects that have certain, relatively stable properties"(2017, p. 754). In her book, Krickel says that entities are concrete physical objects that exist in space and time (Krickel, 2019, p. 117). We also find support for [CKK] in Craver and Darden's 2013 book: "Entities are identified by their properties, by spatio-temporal locations, by boundaries (e.g., a surrounding membrane), by subparts (for example, parts that are chemically bonded to each other and not bonded to parts outside the entity, such as macromolecules), by parts that have their own integrity (for example, organisms as parts of populations), and by their durations (e.g. stable over generations, hours, minutes, or milliseconds.)" (Darden and Craver, 2013, p. 17); this list is probably intended as a disjunction of items that may identify entities, but the passage seems to suggests that identifying[36] entity-types is done by appealing to their corporeal properties.

However, these passages could also be interpreted as applying to a collection of particular entities, and not necessarily to entity-types. Could MDC be nominalist? That is, could MDC be committed to the view that only particulars exist, and types or kinds are nothing other than abstractions where certain details have been omitted?. Beate Krickel seems to be sympathetic to the view that there are no entity-types, but only singular property-instances or *tropes*. For instance, she writes: "mechanistic

[36] Note the success verb employed here.

constitution is a singularist notion. It applies to concrete individuals, i.e., EIOs (entity-involving ocurrents). Type-level generalizations are descriptions that summarize explanatorily relevant aspects of these concrete individuals. Mechanistic constitution is the metaphysical grounding of constitutive relevance, i.e., it explains why claims about constitutive relevance are true" (2018, p. 149).

However, the move towards nominalism comes with a cost. To start, it becomes puzzling why we should expect particular entities falling within a category to behave similarly. If the world is just particular objects and tropes, what makes numerically distinct mechanisms behave similarly under certain interventions? Seemingly nothing justifies projections from some cases to others. Moreover, adopting nominalism would entail giving up part of the realist assumptions behind MDC's distinction between how-possibly, how-plausibly and how-actually models. How-possibly models are conjectures about how the mechanism works; how-plausibly ones, conjectures consistent with known constraints on the components of the mechanism; while "how-actually schemas describe real components, activities, and organizational features of the mechanism" (Craver & Darden, 2013, p. 35). If there aren't real entity-types, but only tropes that are subsumed under a general (pragmatic or mind-dependent) taxonomic category or another, then there isn't an actual distinction between how-plausibly and how-actually models in general explanations. Since there is no mind-independent thing that models' parts should correspond to, there is no sense in which a general model could describe the "real" components of a mechanism, more than another general model employing different taxonomic categories that are equally consistent with the data. In other words: an account of

general models' explanatory success by correspondence to the world, like the one MDC are proposing, is incompatible with nominalism. Thus, the move to nominalism would require MDC to give up one of its most important contributions to philosophy of science —an account of explanatory success in mechanistic explanations. The distinction between how-possibly, how-plausibly and how-actually general models would have to go.

The authors of MDC, however, seem committed to there being repeatables. For instance, Machamer speaks of activities as existing "abstract objects" (2004, p. 30), and Craver and Darden define phenomena "as a repeatable type of event or product" (Craver & Darden, 2013, p. 54); "phenomena […] are the stable and repeatable properties or activities that can be detected, produced and manipulated in a variety of experimental arrangements" (Craver & Darden, 2001, p. 122); . So let's discard nominalism and go back to our original question, how are entity-types to be carved? A possible option is to embrace [CKK] —the thesis that the properties we should be sensitive to when carving working entity-types are corporeal, that is, properties having a determinate spatiotemporal location.

Earlier I mentioned that representation-types may be individuated on the basis of their structural properties, their semantic properties, or both. [CKK] entails that at least the representations individuated on the basis of semantic content —call them "semantic types"— *are not* MDC entities. Therefore, they cannot be mechanistic components in explanations of general phenomena.

Let me illustrate this point with an example: embracing [CKK] precludes there being a general mechanistic model of mice's navigating behavior in the Morris' water maze. Representations of "food is in the north-east" constitute a semantic type, and presumably, they are among what causes the mouse to move towards that direction. We need to appeal to that semantic type in order to explain how mice end up going in that direction as opposed to another, even if (as I suppose) that semantic type is instantiated or underlain by some neural events. If this is something the MDC account cannot accommodate, then such account cannot be used to explain rodent navigation behavior.

There are several responses available to the [CKK]-mechanist at this point, and none of them is promising. The first one would be to split the phenomena to the point in which some neuronal state type could be identified with the previously-semantically-individuated representation. For instance, suppose one could identify the normal neurological basis for a representation with the content "north-east", "south-west", and so on —and do the same for each of the possible directions a mouse can take. So, for instance, the north-east region corresponds to a specific place cell that fires selectively when the animal visits that place in its environment. (Here I am assuming that the representational vehicles underlying each of those location representations are different, which research on place cells supports[37]). Suppose we take the relevant entity to be a given neuron (this particular place cell), as opposed to a representation of a place. Then, we would have sacrificed *generality* in our explanations, for what

[37] Although they do not do so on isolation, but as part of a broader circuit for dynamic representation of self-location, which includes also grid cells.

gain? That move would have created new explananda to tackle. Why are subjects systematic in what representations they may hold? (That is, why is a mouse that has the two abovementioned representations also capable of having the (combined) representation of "south-east"?) Why does the same intervention (e.g. changing the food's location during the training period) affect in a similar manner distinct mechanism-types? If mechanists embrace [CKK], they would be subject to the same systematicity and "representation-hungry" challenges that traditionally anti-representationalists have faced, and it does not seem they would have any more success meeting them than their predecessors.

Moreover, it is not clear that this move —finding a type-identity between the physical and the intentional that holds through time and across individual organisms— is even possible. Consider neural adaptation. When one is perceiving a stimulus, there is a certain frequency of action potential in the neurons coding for that stimuli. However, that spike rate changes with time, being higher when one first encounters the stimulus and becoming lower as one adapts to it as time passes (Webster, 2012). Or take neural fluctuation. Due to the stochastic movement of sodium and potassium ions, occasionally a neuron has a spike in activity that is mere "noise". This makes intervals among spikes vary randomly, even as the representation is held constant (Destexhe, 2012). In both of these cases, the corporeal properties of neurons coding for a certain representational content vary throughout time. And, as the case of neural adaptation shows, distinct spike-rates may the same thing at different times —so it seems that type-identity between semantically- and physically- (e.g. neuronally-) individuated types is not possible. Moreover, hippocampal place cells are known to

undergo remapping; when the mouse experiences the same environment in a new context (e.g. the maze is now filled with water), a portion (or even all) of its place cells remap, meaning that their place field changes location, is lost or a new one is gained. Further, mapping of place cells to an environment differs from individual to individual: the location in the brain of a mouse's place cell encoding for place field F may be radically different from the location in the brain of another mouse's place cell encoding for F. So the prospect of being able to identify a given corporeal object or event (such as a place cell) with a semantic type seem quite meager, unless we give up on explanatory generality altogether.

The second move is to give up the idea that cognitive science is in the business of providing explanatory mechanisms, besides a few exceptions (explanations of individual events, and the subset of cases that may be explained without appealing to representations).

I take it that, whatever move they would make, it would not leave MDC in a desirable position. MDC would be restricting the applicability of their mechanistic account for a good number of seemingly-mechanistic explanations in cognitive science. Further, their position would be equivalent to embracing semantic eliminativism —that is, the view that semantic types have no room in a final, complete science. Semantic eliminativism is a tremendously unpopular position in cognitive science, and is not descriptively adequate of the discipline.

Thus, I think the best option is to read or amend the MDC account of mechanisms as rejecting [CKK]. This is what I will do in the next section.

4. Working-entity-hood

In Section 3, I interpreted mechanists' claims that entities are physical, spatiotemporally located objects as involving [CKK]: the thesis that the carving of an entity-kind must be done on the basis of its corporeal properties. This thesis seems to enjoy some initial plausibility as an interpretation of MDC and related views, judging what has been said about entities, and given that corporeal properties are likely to physically determine which activities an individual entity can engage in.

However, there are reasons to abandon [CKK]. In the first place, the abandonment of [CKK] opens the door to hold a pluralism analogous to the one MDC already hold for activities. MDC's causal minimalism or pluralism is the idea that "causation" is an abstract term we use for what are actually different productive activities. On this view, causation is not a unitary sort of relation between things in the world; rather, there are several specific causal relations or productive *activities*, such as binding, pushing, bending, transcribing, regulating, and so on. These aren't reduceable to e.g. activities occurring at the level of fundamental physics. This view considers that it is up to scientists to determine what activities are and what their characteristics are (see Anscombe, 1971; Godfrey-Smith, 2009). If it is the job of science to determine what the relevant activities are, so it seems it is also their job to determine what the relevant entities are, and what are their characteristics. Placing an *a priori* constraint on what characteristics working entities must have (e.g. the constraint that the entity-

kind must be carved by its corporeal properties) seems at odds with letting scientists discover what the components of reality are.

Second, abandoning [CKK] opens the door to using functional, historical and semantic properties as guides to assess whether an object is a member of a certain mechanistic entity-type in the context of a mechanism. Scientists already do so regularly, e.g. by looking at evolutionary history to see what the relevant entity is, particularly in the case of complex, evolved biological systems. Thus, abandoning [CKK] increases the chances that the MDC account is descriptively adequate of explanatory practices in multiple scientific disciplines.

These reasons won't get us very far, however, unless we can make sense of how non-(transparently-)corporeal properties, such as semantic, functional, or historical properties can be relevant for working entity-hood and causal production.

I mentioned before that any object has an infinite number of properties. Presumably, a subset of an object's properties is relevant to its being a working part of a mechanism, while other properties are irrelevant. This distinction crosscuts the distinction between intrinsic and extrinsic properties: intrinsic properties can be relevant to an object's being a proper part of a mechanism-type (such as the chemical bonds composing macromolecules), but so to can extrinsic ones (such as its spatial location with respect to other components of the mechanism).

An assumption mechanists seem to make is that we need to carve out the working entity-types on the basis of what actually engages in causal production. However, this

would be too rough of an approach: first, because it would allow gerrymandered parts to be proper components of a mechanism (like Franklin-Hall's 2006 "quarter-neurons"). Second, because it would ignore scientific practice of looking at e.g. evolutionary history to characterize the relevant entities, as I mentioned above. Third, because some contingent properties of entities, such as their location or current shape, also determine whether they can engage in activities. Thus, there need to be more constraints on what counts as good entity-types.

Thus, I propose a novel account of which properties mechanists should pick out when carving working entity-*types*. They should consider three sorts of properties, the last one of which seems to be especially relevant at high levels of generality. The properties that are relevant when carving working entity-types are the ones playing the following three roles:

1. **Activity-enabling.** These are the properties allowing entities "to engage in certain activities (and hence [have] certain roles) and not in others" (Craver, 2001, p. 60). The Activity-enabling properties are entities' "specific subset of … properties [that] determine the activities in which they are able to engage" (MDC, 2000, p. 6). That is, Activity-enabling properties are properties that allow the entity to perform an activity; for example, *having a certain electric charge* for bonding, and *rigidity* for pushing.

2. **Robustness.** These are the properties making it the case that entities, like other mechanism components, "have a kind of robustness and reality apart from their place within that mechanism" (Glennan, 1996, p. 53). Although entities must be acting in the context of a causally continuous mechanism to be among its working parts, the robustness requirement implies that entities (and hence some of their properties) must be able to exist outside of their role in a given mechanism, and e.g. allow them to participate in other mechanisms (if the set-up or initial conditions obtain). E.g. having a membrane for *cell*, the forces holding together the atoms in a *molecule*.

3. **Individuating**. These are the properties that we would use in order to individuate the kind in general, outside its place in mechanisms of a certain sort. They make good epistemic proxies (fallible, that is, there may be exceptions) to capture entity-tokens that can engage in the relevant activities. This is because they are normally related to some Activity-enabling or Robustness properties. Sometimes the relation is nomological (e.g. number of protons in the nuclei for *chemical elements* is nomologically related to their polarity), others the relation is a structuring cause (Dretske, 1988): Individuative properties of that sort have shaped or structured the process by which token entities of that kind can engage in certain activities. That is, they have structured the process by which token entities of that sort come to have Activity-enabling and Robustness properties. For instance, having a certain ancestor, for *species*; or proper function in a *module*.

Call this the "ARI account": the properties that are relevant for carving working entities in a mechanism-type are the ones that play the Activity-enabling, Robustness and Individuation roles. In discussion of MDC entities, and in discussions of mental causation more generally, it has been overlooked that the properties that play these roles may, but *don't need to, coincide.* Also, not enough attention has been paid to the fact properties playing each role are significant to various degrees depending on the explanatory project at hand. In general, the more abstract and broad in scope our explanation, the more we will be relying on Individuating as opposed to Activity-Enabling and Robustness properties.

In the rest of this section, I will be articulating this account with the help of some cases. I will start discussing MDC's very own example of a neuronal mechanism: the mechanism of action potential. It occurs when there is a temporary and brief shift (from negative to positive) in the neuron's membrane potential caused by ions suddenly flowing in and out of the neuron. Sometimes these action potentials are modulated by some molecules, which we call "neurotransmitters". Let's suppose we are in the process of discovering the general mechanism of action potential. We got to a point where we are quite certain that instances of that general mechanism have as a component tokens of the activity-type of *binding to a neuron's receptor* (which then produces a change in the receptor-channels, leading to a facilitation, regulation or inhibition of action potential, and so on). But we are not quite sure how to carve the entities engaging in that activity-type. How are we to carve the working entity-type?

The ARI account tells us that we should be looking at some (but not all[38]) individual entities' properties. In particular, we should look at the properties that play the Activity-enabling, Robustness and Individuation roles. Importantly, these roles may be played by different, non-overlapping properties of the same individual. And properties playing certain roles may be more important than others, depending on the scope of our explanatory project. For example, in the case in consideration:

- **Activity-enabling.** *Being spatially located in postsynaptic channel receptors* is what (partly) allows neurotransmitters to engage in the activity of binding to a receptor. This spatial position is what allows a neurotransmitter to play the Participation role: if it didn't have that property (namely, being spatially located in postsynaptic channel receptors), the neurotransmitter wouldn't be able to engage in the activity of binding. Note that this property does not play the Robustness nor the Individuation roles: *being spatially located in postsynaptic channel receptors* doesn't make neurotransmitters robust objects, and it is a property some neurotransmitters may never have, so scientists aren't using it to individuate neurotransmitters. However, as a property required for binding, it (partly) determines working entity-ness.

- **Robustness**. The property of *having chemical formula NH2-CH2-COOH* (for glycine) is what makes of it a robust object, one that can participate in different activities in diverse mechanisms if the right conditions obtain. For instance, this chemical formula binds to glycine receptors, but also makes glycine function as a

[38] An entity has infinitely many properties, and both intrinsic and extrinsic ones may be relevant for their working entity-hood, as I discussed in section 3.

bidentate ligand for metal ions, and a Ph-preserver in some solutions. Chemical composition makes something a robust object and also allows it to participate in activities (since the sort of causal relations or activities it can engage in will be determined by chemical composition). However, given the generality of our target mechanism, this sort of property won't be very useful to delineate the entity-kind.

- **Individuation**. *Having been synthesized in the neuron* is one property neuroscientists employ to individuate and distinguish neurotransmitters from other molecules, including molecules with the same chemical composition that play other biological functions. This property helps identify the entity-kind in general, restricting it to some (but not other) mechanisms. *Having been synthesized in the neuron* is a structuring cause for a neurotransmitter's location: it explains why a significant number of neurotransmitters may come to have the Activity-enabling property of *being spatially located in postsynaptic channel receptors*. This property plays neither Activity-enabling nor Robustness roles. It is not directly responsible for the neurotransmitter to engage in binding (what allows something to participate in binding is not a fact about where it has been synthesized), and neither is it what makes them Robust (since facts about their origin aren't what allows them to engage in these different causal relations). Properties that play the Individuation role may suffice to determine whether something is an instance of a scientifically relevant taxonomical category, but not whether it is a working entity in a mechanism. For example, some neurotransmitters may fail to reach receptors and thus they won't be working entities in the mechanism of action potential.

The neurotransmitter case, I hope, illustrates what the ARI account amounts to when it comes to demarcating which properties matter for working entity-ness. An individual entity needs to have properties of the three sorts in order to fall into the working entity-type for the mechanism of action potential. Some of its properties may just play one such role (as the ones discussed above), others may play several roles (e.g., *having a certain number of electrons* plays Activity-enabling as well as Robustness roles[39]; *having been released by an active neuron* plays Individuative as well as Activity-enabling-roles[40]). Individuation properties will be especially relevant when carving the types for *general* mechanisms, given their relation to Activity-enabling and Robustness ones. This will especially be the case when there is a high quantity of Activity-enabling properties an individual entity needs to possess to engage in the activity, or when there is a lot of variation in which Activity-enabling or Robustness properties tokens possess. For example: in the case of the mechanism of action potential, we can count on Individuative properties being instantiated in virtually[41] every instance of entities engaging in *binding to a neuron's receptor* in the context of the mechanism of action potential. Other properties, however, will predictably vary (e.g. having a certain chemical composition, charge configuration or size properties). In fact, Individuative properties may turn out to be the only ones we cannot abstract away from when we aim to provide mechanistic schemas with a wide degree of generality or scope.

[39] Allowing a neurotransmitter to *bind*, and being a constituent part of the neurotransmitter.
[40] Is what allows neurons to interact in synapses, and distinguishes neurotransmitters from e.g. other amino acids.
[41] Virtually all, but maybe not all. For example, lab-produced, synthetic neurotransmitters may also engage in *binding to a neuron's receptor* in certain circumstances.

The ARI account allows MDC-style approaches to take into account functional and representational properties when carving the relevant entity-types. Let me illustrate this, first, with a toy example that may be familiar to philosophy of mind readers: language of thought (LOT). The mechanism-type of interest is the one behind composition of complex thoughts in Mentalese. That mechanism includes basic symbol-types as a components. Let's say, for our purposes, we are interested in carving the basic symbol-type for the concept we folk-psychologically denominate "EXPENSIVE". ARI would look at the following properties to individuate the basic symbol-type:

- **Activity-enabling.** These will include the vehicle's property of *having logical form of a "predicate"*. Such property enables basic symbols of that type to compose into complex sentential structures during the *combination* activity, which is the one they engage in the context of the mechanism of thought composition (e.g. they combine playing "predicate" role, as opposed to the "relation" one). E.g. the property of *logical form of a Predicate* would allow it to combine with a symbol with the logical form of "entity" such as CAR, producing CAREXPENSIVE (a complex thought with logical form Pa), but not with another symbol with the logical form of "predicate" such as DARK as in EXPENSIVEDARK (logical form PQ).
- **Robustness.** Here we will find properties that are activity-enabling in the context of mechanisms *other than* the mechanism we are interested in. That is, Robustness properties of the basic symbol-type are those that allow it to engage in activities in the context of other mechanisms (that is, mechanism-types that aren't the mechanism for thought composition). For example, *being able to create affective expectations* will be

a Robustness property for the working entity-type of the basic symbol EXPENSIVE. Suppose tokenings of the concept EXPENSIVE produced in response to a product (e.g. wine) have the property of *being able to create affective expectations*. And, in fact, when certain starting conditions obtain, they engage in the activity "creating affective expectations", which in turn impact the product's affective appraisal, resulting in a more enjoyable hedonic experience (something like seems plausible given Plassman et al.'s 2008 wine study, as well as growing literature on how expectations modulate affect). The property of *being able to create affective expectations* is irrelevant for thought composition, but reassures us that the concept EXPENSIVE has a reality outside the mechanism-type of thought composition.

- **Individuation.** Here we will find the semantic content of the symbol EXPENSIVE; that is, the property *being about expensive stuff*. This property is structurally related to Activity-enabling (A) and Robustness (R) ones. It is because a symbol *means expensive*, that it can engage in certain activities and not others in the context of thought composition. The symbol's meaning is a structuring cause of Mentalese combinatorial syntax, which respects semantics. The semantic properties of EXPENSIVE made it the case that entities of that type can combine with symbols of a certain logical form, but not others[42]. Being about expensive stuff is also a structuring cause for its robustness: what makes the case that tokenings of EXPENSIVE create affective expectations is that expensiveness often tracks value. Most likely, that symbol-type's content is also a structuring cause for other properties playing A and R roles not discussed here. That is why, ceteris paribus, the more

[42] This view is compatible with a number of proposals to naturalize semantics.

general[43] the working entity-type, the more we will be relying on Individuation (I) properties to carve it, as opposed to A and R ones —even to the point where we may either abstract away, or ignore, A and R properties.

The one above was a toy example to illustrate how the ARI account works for representations. LOT is somewhat controversial, it is questionable that the concept EXPENSIVE is a basic symbol in Mentalese, and *having the logical form of predicate* might turn out not to be a sort of property that basic symbols in the mind/brain instantiate. The factuality of the example above should not concern us: its whole purpose was to elucidate the ARI account I am proposing, when it comes to distinctively cognitive kinds.

I will discuss another example next, with two goals in mind. First, to show that the ARI account works as well for carving plausible working entity-types —that is, working entity-types for which we have evidence, and that cohere with general cognitive scientific knowledge. Second, and most importantly, that the ARI account is not a substantive amendment to MDC's original (2000) account of mechanisms. Rather, I consider it a *clarification* of what MDC's view entails when it comes to carving mechanism-*types*.

[43] That is, the more we "telescope", to use Darden's term. I will discuss in detail what I mean by this in the next example.

This time, we will be considering the placebo effect for pain. The simplest explanation for the cognitive version of the placebo effect[44] is that, despite beginning with different cues (the effectiveness of an intervention being praised by peers, reading the purpose of a clinical trial in the consent forms, suggestions about the intervention's benefit, etc,), these cues all converge in one representation with the content "this won't be so bad" (or a content along those lines). It is this representation which engages in activities, producing a reduction in the affective component of pain (what makes it "painful" or aversive). The ARI account would carve the working entity-type by taking into account the following properties:

- **Activity-enabling.** This will include properties such as *being connected to evaluative processes, being in a position to impact attention*[45]*,* or *being received as input by evaluative-appraisal systems,* as well as the neuronal properties on which those are grounded, such as *having a certain configuration of some synaptic connections in certain parts of the brain*[46], or *having a certain membrane potential*. Activity-enabling properties are properties of the representation's vehicle that allow tokens of that representation to engage in activities for causal production. Note that, in the case of distinctly cognitive mechanisms (as opposed to merely physiological ones), representational vehicles most often will need to be captured at the algorithmic or

[44] The cognitive version of the placebo effect is mediated by expectations due to e.g. verbal suggestions or information received. In contrast, placebo effects resulting from conditioning work by associating a stimulus with a response. For discussion of the relation between the two, see Montgomery and Kirsch (1997), Colloca and Miller (2011).
[45] These two should be understood as referring primarily to a token's position in the functional organization of the brain, rather than as a capacity or unrealized potentiality.
[46] Think of this as the physical correlate of encoding stored value.

computational level of description[47]. This is because distinctly cognitive activities — such as "retrieving context information" or "modulating affective stimuli appraisals", which occur in cases of the placebo effect— also occur at that level.

To reiterate: I suggest that activity-enabling properties (of entities explaining cognitive phenomena) will often be describable at the algorithmic or computational levels. But, one may wonder, why would that be? It is because activities are "identifiable independently of the individual entities that are acting" (Machamer, 2004, p. 32), based on standard disciplinary knowledge and experimental evidence (Darden et al., 2018, p. 19). Insofar cognitive scientists regularly identify cognitive activities at the algorithmic or computational level, the identification of an entity's activity-enabling properties will likely be at the same level. This is no different from what occurs in many biological mechanism schemas. For activities described at a high level of abstraction and generality (e.g. *expression)*, we may find activity-enabling properties at that same level (e.g. *containing instructions)*. If we go down a level of abstraction and generality for the activity, we will typically do so for activity-enabling properties as well, since they are tightly connected (e.g. *transcription* and

[47] Here I am following Marr (1982) on levels. According to Marr, to understand any information processing system completely, it must be described at three different levels of analysis. These levels are: (i) Computational: the system must be described in terms of the information-processing task it performs. (ii) The algorithmic level: The system must be described in terms of the procedure it uses to perform that task. (iii) The implementational level: The system must be described in terms of the way in which that procedure is implemented in its wetware or hardware (i.e. in terms of the physical properties and processes used by the system to realize the procedure) (Dawson, 1998, p. 288). "According to the tri-level hypothesis, a cognitive system has reality at (and so requires description at) three levels" (Michaelian 2011: 174).

having a promoter, and *translation* and *having a start codon).* This will be the case for multiply realized types in biology, cognition and other special sciences. Marr's three levels provide a system of classification, organized by research questions[48], for what MDC call "degrees of detail" (that is, how many specific details have been dropped or abstracted from) and "scope width" (that is, the size of the domain to which the schema applies, quantified over not by its numerical instances but by the number of explanatory types it can be divided in turn[49]). Marr's tri level hypothesis applies to types presumed to be multiply realizable (that is, having a one-to-many relation between the types of the level above and the ones below). Multiply realiz*ed* types would often vary in detail and scope on a pair: the more detailed a component-type, the narrower its scope. The more abstract, the more general it is. Of course, whether a type is indeed multiply realized is an empirical question, both in biology and in cognitive science. If it is not, degree of detail and scope width will come apart. But for those that are multiply realized, the description of the activity-type and of the activity-enabling property will likely be at the same level.

- **Robustness**. Properties that play the Robustness role are those that make the entity exist outside its place in a particular sort of mechanism. In the case under consideration, the properties playing Robustness roles would be properties that are Activity-enabling in the context of a different mechanism type. For example, our

[48] E.g. "what steps are being carried out to solve the problem?"; "what are the physical characteristics of the system?".
[49] This formulation captures both senses of scope discussed in Darden (1996), organism- (or species)-scope and below-organismic scope (illustrated by the phrase "same cause, same effect"). I take quantification over numerical instances to be unfeasible, as it demands knowing the future as well as the past.

representation has among its Robustness properties *being in a position to influence motor commands* (such as those involved in running away or sighing), and (when combined with other mental states), *being in a position to induce gratitude*. While these Robustness properties may not be specially important within the context of the placebo mechanism, they ensure the entity has independent reality.

Some entities discussed by MDC have natural boundaries: for example, the membranes surrounding a cell's nucleus (Darden, 2008, p. 960). Natural boundaries, if they exist, help us identify entities. But in fact, it is not necessary that entities have natural boundaries to be robust. All that is needed is that they "have a kind of robustness and reality apart from their place within that mechanism" (Glennan, 1996, p. 53). But what does that mean? Glennan tells us that "it should in principle be possible to take the part out of the mechanism and consider its properties in another context" (Glennan, 1996, p. 53).This is perhaps too strong, as "some parts of a mechanism might become unstable when removed from their mechanistic context" (Craver & Tabery, 2019). I don't think we should take Glennan's suggestion literally, but metaphorically. Observing the entity acting and being a producer of change in the context of another mechanism (as when it modulates motor commands, in the context of the mechanism for sighing) already shows that it can play a causal role outside of the placebo effect mechanism. It indicates its reality in another mechanism-type context. In other words: assuming (like MDC do) that what engages in causal production is the token entity itself (as opposed to its properties), it becomes a perspectival matter which of its properties are the Robustness ones and which are the Activity-enabling ones. It will depend on which mechanism-type we are interested in.

Are we interested in explaining sighing? Then, *being in a position to influence motor commands* is an Activity-enabling property. Are we interested in explaining pain reduction? Then, it is a Robustness one. That is why I speak of properties playing certain roles (as opposed to being different kinds of properties in some strong metaphysical sense).

- **Individuation**. *Having the content "this won't be so bad"* is a property we would use to individuate the kind, within and outside the context of the placebo mechanism. This is because *semantic* properties are often structurally related to Activity-enabling and Robustness ones: it is because Activity-enabling properties have this content, that the cognitive system is made in such a way that it can engage in the relevant activities.

In the neurotransmitter example discussed in the first section, we saw how certain historical properties (i.e. *being synthetized in a neuron*) act as structural causes for Activity-enabling and Robustness ones. Here, I claim that content, too, structures properties playing the two other roles. And I am not alone on that claim. Nicholas Shea (2018) provides a very compelling case for content being explanatorily relevant. According to him, content needs to be appealed to in order to make sense of a cognitive mechanism's configuration and behavior —that is, why the mechanism is set up the way it is, and why it does what it does. There's a similarity between Dretske's discussion of structuring causes and Shea's view that content has explanatory purchase: both treat content as what explains the item's role in producing (proximally) certain phenomena. In Shea's view, evolution produces organisms

which robustly produce certain outputs. By having an entity that keeps track of aspects of the environment (a representation with certain content), that information can be mapped to certain outcome that has been the target of stabilization[50]. So representations aren't explanatory fictions (or a mere "gloss" on non-semantic explanations), but very real entities that arise when stabilization and robustness are achieved by their vehicles bearing exploitable relations to distal things in the world, and reliably co-occur with it. Because of this, overlooking content-properties may result in a failure to capture vehicle-world relationship patterns that would otherwise have been missed. We miss explanatory generality if we individuate the entities based on their non-semantic properties. So the content *"this won't be so bad"*, even if it does not engage in activities directly, needs to be appealed to for individuation, since it allows us to capture relevant general patterns, and is connected to (and therefore serves as a proxy for) Robustness and Activity-enabling properties — explaining why these came to play the roles they do within a mechanism, as opposed to another.

5. Getting it ARIght

In the previous section, I introduced, illustrated and articulated the ARI account for carving working entity-types in the context of general mechanisms. I consider ARI to

[50] Stabilizing processes, for Shea (2008), include (i) natural selection, (ii) learning, and (iii) contributing to the organism's persistence.

be an extension, rather than an amendment, of the account of mechanisms discussed in MDC's (2000) paper. Nevertheless, whether it is an extension or an amendment, in this section I will argue that my approach aids progress in constructing general mechanistic explanations inside and outside cognitive science.

First, a virtue of the ARI analysis is that it makes sense of certain practices in the special sciences. It accounts for why, e.g., properties relating to an item's ancestry or evolutionary history may help us identify what are the working parts of the mechanism, even when such properties are not capable of proximally initiating causal production. Epistemically, properties playing Individuation roles serve as a good proxy to capture individual entities that likely will be capable of engaging in the relevant productive activity. For example, facts about ancestry (which some philosophers of biology take to determine one's species) may serve as a good (albeit fallible) proxy to capture creatures that have the mammary glands necessary to *nurse* (an activity) their young. Metaphysically, there is a grounding relation between properties playing Individuation roles, with those playing Activity-enabling and Robustness ones. Sometimes, this relation is nomological (e.g. number of protons in the nuclei for *chemical elements* is nomologically related to their polarity). In cases involving living evolved organisms, this relation will often be structural: Individuative properties would have shaped or structured the process by which entities of that sort can engage in that activity-kind. For example, facts about ancestry shape whether creatures have mammary glands, and therefore are able to engage in *nursing*.

Second, the fact that properties playing roles A, R and I do not need to overlap has interesting consequences for the MDC account of mechanisms. To start, it explains away Franklin-Hall's (2016) "carving problem" or "quarter-neuron problem". Franklin-Hall's carving problem for mechanisms is the problem of how to carve a mechanism's components in a way that is appropriate, non-arbitrary and non-gerrymandered. Insofar as Franklin-Hall is demanding an in-principle rule to carve entity-types, if minimal entity pluralism is the case, there is no way to meet that demand: different entity-types may require different ways of carving, and there may be no generalizations that apply to all of them. However, insofar Franklin-Hall is demanding a principled way to distinguish between good component parts (like neurons) from bad component parts (like quarter-neurons), there is something the mechanist could say: that good component parts have properties playing not only Activity-enabling, but also Robustness and Individuation roles. In my opinion, the "quarter-neuron" problem's appeal[51] derives from conflating Activity-enabling with Individuation properties. If one thinks that components of a mechanism-type must be individuated only by *what allows it to engage in the activity*, then we get gerrymandered components —at least, in cases where *what allows it to engage in the activity* is spatially localized in a segment (but not all) of the token entities. Franklin-Hall seems to suggest that working entity-types can be individuated using just abstraction over activity-enabling properties. But that is just not the case. The sorts of properties we should necessarily look at to individuate the type, particularly when

[51] Given that, if we "follow the science" (as we should), there may not be a universal *a priori* way to distinguish "good" from "bad" entity-types.

dealing with general (that is, with some scope breadth) mechanisms, are those that are structurally or nomologically related to Activity-enabling and Robustness ones (i.e. Individuating properties). Abstracting over Activity-enabling properties won't usually give us a type so related to the presence of the other two. If one carves the working entity types taking into account Individuation properties, as well as Robustness and Activity-enabling ones, one does not end up with gerrymandered components. A different way to put the same point is this: a principled way to distinguish good from bad entity-types in a general (mid- to wide- scope) mechanism model is by observing whether hypothesized working-entity types track Individuation properties.

Lastly, ARI preserves our intuitions that entities as a whole (as opposed to only parts or only properties of them) act as causes. Consistent with the MDC account, it assumes that "an entity acts as a cause when it engages in a productive activity" (2000, p. 6), and productive activities do not always occur at the fundamental level (as entailed by minimalist causal pluralism). I do not think causal exclusion worries (of the sort discussed in Kim, 1988, 1989; Shapiro & Sober, 2007) threaten my account in particular. First, mechanists already talk about non-fundamental (maybe weakly emergent?) entities engaging in causes (activities). They say, e.g., that neurons and neurotransmitters are entities engaging in causal activities in the context of the mechanism of action potential (Machamer et al., 2000), that a virus produces symptoms as part of a disease mechanism (Craver & Tabery, 201), that a heart *pumps* blood, curare *enters a creature's bloodstream*, and so on (Darden and Craver, 2013). Neither neurons, neurotransmitters, hearts or viruses are fundamental (basic) physical objects. Thus, causal exclusion worries are applicable to the entire project of

mechanistic explanation, not just mechanistic explanation of cognition; and dealing with them is a whole separate project beyond the scope of this paper. Second, MDC has a minimalist pluralist view of causation, according to which the sorts of causation there are is something to be determined scientifically, not from the armchair. Thus, in their view, there should not be an impediment in principle to science positing a special sort of causation (an activity-type) that (only) representations can engage in. If causal exclusion worries are something MDC has to deal with in general, they don't pose a worry in particular for my proposal, so I don't think it's incumbent on me to sort them out. I am not making MDC's position subject to more problems than the ones they already face.

To summarize: the *properties (plural)* that are relevant to working entity-ness in the context of a general mechanism for a phenomenon are those that play [1] Activity-enabling roles; [2] Robustness roles; [3] Individuation roles. Having properties playing these three roles is what makes of something a working entity in the context of a mechanism. A given property P may play certain roles and not others. When it comes to general explanations involving kinds, there is nothing problematic in individuating kinds by virtue of their semantic properties (which are Individuating properties), even when those are not Activity-enabling ones, and therefore do not play a direct[52] causal role in the workings of a mechanism. A nomological or structural

[52] The role, however, may be indirect or structural, that is, having a contributing or causal role to the triggering direct cause to produce its effect (as in Dretske, 2004).

relation between Individuative and Activity-enabling & Robustness properties is all we need to get general explanations off the ground.

6. Conclusion

This paper began with the idea that [general explanations of cognitive phenomena involving representations in the explanans] may not have a place in MDC's account of mechanisms. However, despite some suggestions to the contrary, I argued that there is such a place in the account for representations, once we get a clearer sense of what carving a general mechanism's working entity-types involves. That clearer sense is provided by my ARI account.

Despite what may have at first seemed to be the case, the MDC account of mechanisms can include representations as explanans. The relevance of the New Mechanistic philosophy for cognitive science is thus vindicated.

Chapter 4: A material girl in a normative world: the extreme reasons-irresponsiveness account of mental disorders

1. Introduction

One of the tasks of psychiatry is to classify and taxonomize mental disorders (I will be using that term interchangeably with "psychiatric disorders", "mental illnesses", and "psychopathologies" throughout the paper). Mental disorders include pathologies such as schizophrenia, depression and autism spectrum disorder. There is a plurality of classification systems in psychiatry: for example, the Diagnostic and Statistical Manual of Mental Disorders (DSM), the Research Domain Criteria (RDoC), the Hierarchical Taxonomy Of Psychopathology (HiTOP), and the World Health Organization's International Classification of Diseases (ICD) all provide alternative systems of classification for psychiatric disorders. In this paper, I will treat "mental disorders" as co-extensive with "DSM diagnostic categories"[53], as in the Western

[53] The DSM undergoes periodic change on its classifications. As of 2013, the American Psychiatric Association has moved to an Arabic numeral notation for the different versions of the DSM, suggesting that changes between versions of the DSM may appear more frequently and in a more piecemeal fashion than they did before. While we can treat the book by the American Psychiatric Association (2013) (henceforth: DSM-5) as the best psychiatric classification system currently available given its widespread use by mental health care providers, that doesn't mean it is final. It is to be expected that in some years from now some diagnostic categories will disappear, others will be merged, and some will be added. Thus, when discussing what sort of thing mental disorders are, I do not intend for my claims to be tied to the particular current list of psychopathologies in the DSM-5, but rather, to be general claims about mental disorders presently included in the DSM or that we can foresee progressive versions of the DSM will include.

context (or at least within the United States), the categories of the DSM are the most common way mental health providers, researchers and the general public think of mental disorders —and these categories are intended to be applicable worldwide.

Part of the philosophy of psychiatry literature has been trying to advance an answer to the question: what are mental disorders? Having an extensional characterization of mental disorders, as I did above, does not tell us what they *are*. Although this is a foundational question for psychiatry, and one critically important for patients and clinicians, it has resisted a straightforward answer. Further, this question can be decomposed intro three more nuanced ones: first, what *sort* of metaphysical nature do mental disorders have? Are they natural kinds, socially constructed kinds, broken normal kinds, pragmatic kinds…? Second, what makes a mental disorder a *disorder*, rather than a non-pathological feature of a person? Third, what makes a disorder *mental*, as opposed to a neuronal or otherwise somatic condition?

These questions are interrelated, although the difference in emphasis requires us to bring different considerations into account. In this paper, I will explain why I think the most promising question to pose in order to understand the nature of mental disorders is the third one, and I will provide an account of what makes disorders *mental* that will have implications for the other two questions. In particular, I will be defending the claim that mental disorders are alterations affecting the *reasons-responsiveness* of certain mental states and behaviors —those for which we have evaluative normative standards.

To do so, I will first argue that the two first questions aren't a promising venue to get at the nature of mental disorders. In section 3, I will present some answers to the question of what makes certain pathologies *mental*, before introducing my own account. In section 4, I will present some considerations supporting my view. Chapter 5 attempts a historical explanation for the proposed regulative ideal behind psychiatry, before concluding in section 6.

2. Two questions about mental disorders

What are mental disorders? The DSM-5 characterizes them as follows:

" A mental disorder is a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning. Mental disorders are usually associated with significant distress or disability in social, occupational, or other important activities. An expectable or culturally approved response to a common stressor or loss, such as the death of a loved one, is not a mental disorder. Socially deviant behavior (e.g., political, religious, or sexual) and conflicts that are primarily between the individual and society are not mental disorders unless the deviance or conflict results from a dysfunction in the individual, as described above. The diagnosis of a mental disorder should have clinical utility: it should help clinicians to determine prognosis, treatment plans, and potential treatment outcomes for their patients. […] It should be noted that

the definition of mental disorder was developed for clinical, public health, and research purposes. Additional information is usually required beyond that contained in the DSM-5 diagnostic criteria in order to make legal judgments on such issues as criminal responsibility, eligibility for disability compensation, and competency" (DSM-5 Text Revision)

Two important remarks from this. The DSM uses operational definitions for disorders —that is, disorders are characterized by their symptoms, as opposed to their etiology. For example, a person has e.g. depression if she doesn't meet the exclusionary criteria and shows 5 out of 9 symptoms of a list —symptoms that are meant to be sensitive (able to detect depression) and specific (able to distinguish those who are depressed from those who aren't).

Also, the taxonomy of mental disorders is tailored to suit a particular purpose within the context of clinical practice and research. They are not (to be) assessed, thus, by the same standards we would judge the taxonomies of other sciences.

These two features, I think, make it unlikely that psychopathologies conform to a homogenous class, or have some sort of metaphysical nature in common. Despite this, some authors propose blanket answers to the question: what sort of thing are mental disorders? They have been said to be natural kinds (Tsou, 2022*)*, harmful failures of functions that were selected for (Wakefield, 1992), mechanistic property cluster kinds (Kendler et al., 2011), networks of causally connected symptoms (Borsboom, 2017), and extreme points in a continuum of functioning (Cuthbert and Insel, 2013; Kendell, 1991). Others (e.g. Foucault's 1975 discussion of "madness";

Hacking, 1999) have argued that mental disorders are social constructions, in the sense that they are determined or controlled by psychological and sociocultural factors.

In my opinion, it is unlikely for the different psychopathologies to be deeply similar in nature. Psychiatry is an "applied", as opposed to a "pure", science. The aim of pure sciences is merely descriptive and explanatory, or to get true characterizations of the world[54]. Applied sciences, in contrast, also have built-in a *goal* or something regarded as valuable that they seek to promote[55]. In the applied sciences, the selection of phenomena of interest responds *not purely* to facts[56] about how the world is put together (including facts about whether something is socially constructed or whether there is variation in a continuum). Rather, what falls under the subject matter of an applied discipline depends on *which facts are such that, if we knew them, would be relevant in advancing the pursuit or promotion of the goal in question*. Thus, their subject matters (and their resulting taxonomies) may cross-cut natural domains of things. For example, bladder cancer risk is the result of the interaction between a genetic factor —variation of the NAT2 gene— and an environmental factor — smoking behavior. The same seems to be the case for psychiatry: for instance, eating disorders have been said to result from genetic as well as environmental factors

[54] Obviously, there may be applications or uses to be developed from the knowledge about the world that we gained, but these usually fall outside of the science itself, as a matter for perhaps engineers, inventors or practitioners.
[55] Whether all sciences are value-laden is a claim I won't explore here; for our purposes, it suffices to say that even if all sciences are value-laden, such value-ladenness admits of different degrees, such that physics and chemistry are less value-laded than e.g. economics and animal welfare research; and my claims can be reconstructed as saying that psychiatry as a discipline would fall on the higher end of the spectrum of value-ladenness.
[56] E.g. facts about similarity in some phenomena, or the things that explain them.

(Mazzeo & Bulik, 2009), and exposure to poverty and violence plays a role in the onset of major depression (Satcher, 2001). It should thus be unsurprising that the pathologies medicine deals with don't have a common underlying metaphysical nature. Some medical conditions are plausibly understood as extreme points in a continuum of natural variation on functioning (e.g. blood having a reduced ability to clot, for hemophilia); others seem to be broken normals (e.g. broken bones, for fractures); some seem to be natural kinds (e.g. cases of AIDS involve a common etiological mechanism involving the HIV virus); some are mere umbrella terms for a given phenomenon independently of what caused it (e.g. acute myocardial ischemia is a serious reduction in blood flow to the heart for whatever cause). Like medicine, psychiatry's taxonomical categories plausibly may pick up different sorts of things. For instance, type B personality disorders have been said to be moral kinds (Charland, 2004, 2006), depression a mechanistic property cluster kind (Kendler et al., 2011), dyslexia an evolutionary adaptation (Garson, 2022b), homosexuality a socially constructed kind (Szasz, 1960), paraphilias socially constructed interactive kinds (Hacking, 1999), and so on. Moreover, it is possible that at our stage of understanding, one and the same mental disorder may be understood in different ways, depending on the metaphysical theory one favors. For instance, autism spectrum disorder may be seen as a point in a continuum of functioning and neurodevelopment (Chawner & Owen, 2022); it can also be seen a breakdown of the mind-reading system (Baron-Cohen, 1995); as a pathology with biological etiology related to abnormal gut microbiome (Kang et al., 2019); or as an umbrella term for what are actually different conditions (Whitehouse & Stanley, 2013). My point here is

not to favor a given theory of autism; rather, to point out that different mental disorders may be different things metaphysically, and that is consistent with them being mental disorders.

That psychiatry is an applied science does not mean that the discipline isn't (or cannot be) scientifically rigorous: evidence-based medicine follows high standards of scientific rigor; so does behavioral research conducted by employees of Booking.com that aims to increase number of hotel bookings made through the website. But having a practical goal impacts how to judge a discipline's success and progress. Pure sciences are progressing when they approximate or better capture the way the world is[57], independently of whether these truths are useful to us or fulfill any goal (other than understanding the world itself). In contrast, applied sciences' progress is to be judged by whether they generate knowledge that *advances the pursuit of the goal* in question. If a clinical trial revealed facts about human physiology that didn't help in advancing a path for treatment of the condition in question, the clinical trial would have failed. If Booking.com research captured true generalities about human psychology that could not be used to further its revenue-increasing goals, they would have failed. Similarly for psychiatry: diagnostic categories get decided in the context of promoting mental health. It shouldn't be surprising, thus, that the taxonomical categories applied sciences employ are not the language of nature itself.

[57] Here, I am using a language that is neutral regarding the scientific realism vs anti-realism debate. According to Fine (2001), both scientific realism and anti-realism hold that there are things like atoms, or that our best theories are approximately true. The difference is that realists use a correspondence theory of truth, according to which a scientific theory is true if it corresponds to the underlying structure of the world, while anti-realists hold an alternative theory of truth, like a pragmatist or evidential one.

Does that mean that mental disorders are purely pragmatic tools, or practical kinds, as Zachar (2002, 2014, 2015) has argued? Zachar holds the view that mental disorder categories (our concepts of psychopathologies) are the products of the clinical goals of practitioners and patients, commercial interests, the scientific goals of researchers, priorities of health care administrators, and so on (Zachar, 2015, p. 290). This does not mean that mental disorders aren't objectively grounded in features of the causal structure of the mind/brain: they are so grounded, but they are not reducible to these (Kendler et al., 2011). While I tend to agree with Zachar (for many psychopathologies, at least), I think his theory is in a certain respect uninformative — in that it does not tell us what serves as a *regulative ideal* for mental disorders, i.e., it does not tell us what heuristics are being used to consider something as a mental disorder, or what the limits or constraints are on what can acceptably (for the parties involved) be included in that notion.

Thus, we may change the emphasis of our question and wonder instead what makes mental disorders *disorders* —as opposed to non-pathological features of persons. Maybe this question is a more informative one to ask, compared to the question of what is the metaphysical nature of psychopathologies.

An answer that emphasizes the *disorder* part of psychopathologies will link them to the medical model, and typically emphasize one of three points: (i) that mental disorders are "usually associated with significant distress or disability in social, occupational, or other important activities" (DSM-5, Introduction), which are objectively bad; (ii) that mental disorders "involve departures from normal

functioning of a biological system" (Murphy, 2006, p.44), or (iii) that mental disorders can be medically treated. In my opinion, none of these will give us a complete picture of what psychopathologies are, as I will briefly explain next.

While experienced distress and/or disability seem to be important aspects of what makes a certain condition pathological, they seem to be neither necessary nor sufficient. It is well-known that some conditions in the DSM do not seem to involve either, such as manic episodes and narcissistic personality disorder. Sometimes, socially or culturally deviant behavior (e.g, inappropriate staring, a propensity to gossip about people behind their back, certain ways of dressing) may result in disability and distress (e.g. social condemnation and exclusion) without there being anything pathological about the individual. Further, distress seems to be a normal feature of human life: most of us would experience distress e.g. if we were to incur a significant financial loss, or experience an unwanted breakup. There's also some concern that making disorders depend on a self-assessment of what is acceptable suffering makes psychopathologies subjective.

Some philosophers have instead proposed that mental disorders are biological dysfunctions involving the mind/brain (e.g. Wakefield, 2007; Szasz, 1960). If indeed mental illnesses were the result of neurophysiological breakdowns, mental disorders would be objective and independent of value judgements. However, there are reasons to be wary of this sort of approach. First, it makes the discernibility of mental disorders dependent on whether the proper function of mental components is known. A taxonomical categorization of psychopathologies would require knowledge of the

different components of the mind/brain, as well as what their proper function is. Even if that could be obtained, it is definitely not the case that current categories of psychopathologies reflect these. Second, this approach is at odds with the DSM's commitment to provide operational definitions only, as opposed to hypotheses about their origin. Such an approach would not capture how psychiatry currently thinks about mental disorders, and thus will likely be descriptively inadequate. Third, there is a plausible case to be made that some psychopathologies are neither a disease or a defect, but a designed feature (Garson, 2022a). Fourth, one may worry that such a characterization would ultimately erase the distinction between mental and somatic disorders (Bolton & Hill, 2004).

Lastly, one might want to make the notion of disorder dependent on that of treatment. For instance, the fact that lithium is an effective treatment for mania, and selective serotonin reuptake inhibitors are an effective treatment for depression, seems to speak to their being neurophysiological pathologies. Pickard (2009), for example, suggests that Type B personality disorders are indeed psychopathologies because their symptoms can be improved using antidepressants and psychotherapy. However, if we take as our criterion for something to be a psychopathology to involve a series of negative symptoms that tend to cluster together and improve by a specific treatment, we will include many things that aren't psychopathologies as such. For example, being stressed, anxious, impulsive, and lacking emotional balance tend to cluster together for many Western individuals, given chronic lack of sleep (Suchecki et al., 2012; Goldstein & Walker, 2014; Simon et al., 2020). It can, however, be "treated" (I am assuming that recommendations or prescriptions given by a clinical practitioner

with the purpose of remedying the patient's symptoms count as therapy) by getting regularly 8+ hours of sleep at night. Sleep deprivation does not seem to be a mental disorder. Improvement of symptoms as a result of medical treatment does not necessarily imply that the patient was experiencing a psychopathology. For example, antidepressants may be effective in treating the symptoms of non-pathological bereavement responses, and ADHD medication is commonly used by regular students to improve concentration. This account of mental disorders as clusters of symptoms responsive to treatments is too encompassing.

While my remarks in this section are by no means definitive, I hope they have served to motivate the view that something else is needed in order to capture what's uniquely characteristic of *psycho*pathologies. Thus, I think the question: "what makes a disorder *mental*, as opposed to somatic?" is a more interesting one to pose. I turn to this question next.

3. What makes a disorder *mental*?

In the philosophy of mind, there have been various attempts to find a feature or set of features that all mental states, processes, properties and so on have, and that all non-mental stuff lacks. These feature(s) are sometimes called "the mark(s) of the mental" or criteria for mentality. The challenge is to provide such "mark(s)" in such a way that does not immediately presuppose the truth of a particular stance on the mind-body problem. There is still an ongoing debate about what the "mark of the mental"

is, but two features seem to be the most common candidates. First, phenomenal consciousness: it is the property which mental states have when it is like something to undergo them (Nagel, 1974). Second, intentionality: the property of being *about* something, of being directed at something, of standing for something (Brentano, 1874). There are also other less popular proposals, such as direct access, incorrigibility, and transparency, but I won't discuss them here. Neither of these common two proposals, however, is specific enough to demarcate mental disorders from other sorts of pathologies. For instance, congenital blindness and prosopagnosia are disorders of perception, blindsight and split-brain patients have abnormal qualia, and aphasias involve abnormal intentionality, but none of these is a mental disorder —despite involving alterations of intentionality and/or phenomenology. In general, it seems that pathologies involving perceptual states (e.g. blindness, tinnitus) aren't psychiatric disorders. But, perception and perceptual states are "mental states", nevertheless.

Another possible way to cash out the domain of the "mental" in psychiatry is by appealing to the faculties of faculty psychology: attention, perception, memory, and so on (Wakefield, 1997; Graham, 2010). However, this does not seem to explain why certain conditions affecting a particular faculty (such as prosopagnosia or blindsight) are not included in the DSM, while others that don't appeal to any system of faculty psychology qualify (e.g. eating disorders, personality disorders, bipolar disorders).

The proposal to characterize the "mental" negatively, that is, by contrast to what we know to be physical or have a physical cause, does not score much better either. On

the one hand, there is no reason to suppose the mental is non-physical in nature. Assuming physicalism is true, everything that is mental has a physical supervenience base, and therefore any mental disorder would also be a disorder with a biological or neurological basis. Psychiatrists don't seem to endorse dualism. Maybe mental disorders are those whose physical cause is unknown? This doesn't seem to be the case either. Some psychopathologies included in the DSM have a known somatic cause: for instance, Down syndrome is the product of having a third copy of the 21st chromosome. Moreover, much psychiatric research tries to understand the physical causes of psychopathologies. If the *mental* of mental disorders referred to that with unknown physical causes, it would then seem that the success of psychiatric research would be to push psychopathologies outside of its domain!

So, on what grounds are some conditions considered mental disorders, while others aren't? In the remainder of this section , I will provide my own account of what makes psychopathologies *mental* disorders. I think it is possible to provide a general characterization of mental disorders that captures the regulative ideal informing the paradigmatic conditions listed in the DSM, drawing on a conception of the mental that doesn't come from philosophy, but from a subset of the "mental" according to folk psychology. Folk psychology is our common-sense theory of mind: it is not a scientific theory built on the basis of evidence (and as such, it is unlikely to be true), but rather a tacit, internalized theory we use to explain and predict people's states and behavior. Folk psychology is a product of the functioning of our mind-reading system in a cultural and linguistic context in which we, during development, acquire different terms for different mental states, rationalizations for people's behavior and states, and

normative standards to evaluate them. What I claim here is that the notion of the "mental" psychiatry relies on is informed by folk psychology. In particular, psychiatry is concerned with the (subset of) mental states and behaviors folk-psychology considers generally explainable by, and responsive to, *reasons* (as opposed to it being explainable merely by appealing to *causes)*.

Excluded from these would be the mental states and behaviors that folk-psychology regards as causal products that aren't reasons-responsive —including perceptions and reflex acts. For instance, seeing a red tomato is not responsive to reasons. Thus, psychopathologies aren't pathologies affecting mental states –but only those mental states and mentally-caused behaviors whose production is reasons-responsive, according to a folk-psychological understanding of the mind. The domain of interest for psychiatry, thus, includes emotions, beliefs, desires, behaviors, moods, worries, intentions, anxieties, and other states and acts whose production changes depending on the reasons available, and is subject to normative standards of appropriateness. What makes all of those (including overt behavior) belong to the domain of the mental is that they are usually produced in a reasons-responsive way, according to our common sense understanding of psychology. We are now in a position to understand psychopathologies as alterations of such modes of production:

> **Mental disorder.** A mental disorder is a pathology affecting how (some) mental states and acts are produced, such that the *degree of reasons-responsiveness* of the mechanism(s) responsible for their production is *significantly diminished*

> with respect to what we would folk-psychologically expect it to be —or, in the most extreme case, the mechanism(s) responsible for their production *is no longer responsive to reasons.*

On this view, psychopathologies are alterations in the reasons-responsiveness of the mechanism(s) producing certain mental states and/or behaviors. Importantly, just having mental states or actions that are unfitting (inappropriate) given the reasons available to the agent does *not* qualify as having a mental disorder. One does not have a mental disorder just because one is terrified of inoffensive dogs, because on the basis of little to no evidence one believes (mistakenly) that one is being followed, or because one tends to get quite angry at mundane stuff. One may end up holding wrong or unfitting beliefs, emotions or actions by a myriad of paths. What matters are the psychological mechanisms that enable one to have these sorts of mental states and behavior. Compared to those of non-disordered persons, are they sensitive and reactive to the appropriate reasons, such that (at least in a good number of cases) if a person has sufficient reasons to the contrary, their output would be different? Do reasons tend to modulate the outputs produced by the mechanism? A negative answer to those questions indicates the presence of a mental disorder.

At this point, I should make a few preliminary clarifications: although I have been talking of "folk psychology", there isn't one folk psychology but several partly overlapping folk-psychologies, since our common sense "theory of mind" is partly acquired via socialization and acculturation. Thus, there will be culture or sub-

cultural differences in, at least, the concepts employed to describe mental states and actions[58]; the generalizations and predictions made over them[59]; and the normative standards pertaining to mental states and/or actions[60]. Furthermore, these may change over time, to the extent that folk psychology, reasons-discourse and normative standards evolve alongside socio-cultural factors. All this implies that disorders too may change over time, and some disorders may be culture-specific, or at least have culture-specific symptoms. For instance, experiences of "being possessed" is a specific symptom for Dissociative Identity Disorder in certain cultures (those where folk-psychology includes possession as an explanatory category). Homosexuality was, and stopped being, a disorder as same-sex romance gained more social acceptance in Western contexts, and it was no longer seen as an unfitting response to individuals of the same sex. In the context of certain cultural and religious practices, avoidance of food intake is not a symptom of Avoidant/Restrictive Food Intake Disorder. The fact that there is a plurality of partly overlapping folk psychologies has consequences for what counts as a disorder. Certain psychopathologies may be universal, while others may not be, or the symptoms may be different across divergent cultural contexts.

[58] For instance, English lacks a term to describe the feeling of deep longing or melancholy that is supposed to be characteristic of Portuguese and Brazilian cultures (*saudade*), which Portuguese writer Manuel de Melo has described as: "a pleasure you suffer, an ailment you enjoy."
[59] For example, in some cultures there is a habit of drinking while having a meal, so the absolute number of alcoholic beverages consumed per week would not give us a clear indication of alcoholism.
[60] E.g. Japanese cultures tend to discourage overt displays of emotions. Being disgusted at the prospective of eating snails would be a fitting reaction in the US, but not in France, where they are considered a delicacy.

I also agree with those who consider that folk psychology is not the correct theory of the mind, and that such a theory must come from cognitive science (e.g. Dennett, 1985, 1987; Carruthers, 2011, Heyes, 2018). Assuming that the distinction between "the mental" vs "non-mental" has some sort of interest-independent reality, using folk psychology we would probably fail to get the division right, given the large amount of evidence that we interpret, and to certain degree confabulate, the reasons behind our actions and those of others (Carruthers, 2011). Thus, it is possible that at least some parts of folk psychology are just useful fictions —including those that consider certain mental states and behaviors as reasons-responsive.

Let's go back to the view for now. In philosophy, reasons-responsiveness has received the most attention in ethics, especially in the context of the debate between compatibilism versus incompatibilism. Contrary to incompatiblists, compatibilists hold the view that causal determinism is compatible with free will and moral responsibility. After Harry Frankfurt's (1969) famous argument that moral responsibility does not require the ability to do otherwise, source or production compatibilists tried to find an account of free action in terms the action's source, of the actual sequence of events leading to the action (Fischer, 1994). Many consider reasons-responsiveness to be the mark of moral responsibility (e.g. Wolf, 1990; Fischer, 1994; Haji, 1998; Wallace, 1996; Nelkin, 2011; Sartorio, 2016). In that spirit, John M. Fischer and Mark Ravizza (1998) provide a very influential account of moral responsibility according to which a person is morally responsible for her behavior if she has guidance control over it, that is, if the mechanism producing these behaviors is *responsive to reasons*. Guidance control does not require access to alternatives:

instead, it requires that the action in question (i) is produced by the agent's own mental mechanisms, and (ii) that such mechanisms are moderately reasons-responsive.

Fischer and Ravizza do not cash out what exactly constitutes a mechanism, or which they are, but that notion is meant to include any psychological mechanism that causally issued in the person's action. By the agent's "own" mechanism, Fischer and Ravizza mean that the process that leads to the behavior has not been e.g. "'implanted' by a scientist or produced via direct electronic stimulation of the brain" (Fischer and Ravizza, 1998, p. 230). Moreover, the condition also requires that the actual sequence mechanism, or psychological mechanism that produced the person's action, is one for which the agent in the past has "taken responsibility". This does not mean that the person is able to correctly identify the mechanism producing her action; that would require possessing a lot of cognitive scientific knowledge that cannot be attributed to most people. Instead, the demand here is that historically, one has seen oneself as the source of such actions as well as the fair target of others' reactions to them, based on evidence. The evidence appealed to here is tightly connected to the second condition. The second condition Fischer and Ravizza lay out is that such a mechanism is moderately reasons-responsive. The requirement here is counterfactual: were the agent in "relevantly similar" scenarios where sufficient moral[61] reasons *to do otherwise* were present, at least some times the mechanism would have produced different outputs. This can in turn be cashed out as the mechanism having two

[61] In their original formulation, Fischer and Ravizza didn't specify that those reasons had to be moral, but that was later added.

characteristics: (i) sensitivity, or capacity to recognize at least a significant range of reasons (including moral reasons) for and against so acting, and (ii) reactivity, or the capacity adjust behavior in accordance with at least some of those reasons (Zimmerman, 2003). Fischer and Ravizza's is a *moderate* version of reasons-responsiveness, as it requires the mechanism to have moderate sensitivity (i.e. is able to recognize a good range of, but not all, reasons to do otherwise) and weak reactivity (i.e. is able to issue a different action in at least one scenario).

Whatever the merits of Fischer and Ravizza's view to capture moral responsibility[62], I think their account can be used as a basis to illustrate the account of reasons-responsive mental states and behaviors that serves as a regulative ideal for psychiatry. This will require some modifications [63], including, at least, expanding it in two fronts.

- **Outputs**. In the case of the "mental" for psychiatry, the relevant outputs of the mechanism won't be *only* actions or overt behaviors: outputs would *also* include moods, emotions, beliefs, thoughts, doubts, worries, desires, and other mental states whose production ordinary discourse considers to be reasons-responsive, and regards these outputs themselves as *fitting* or *unfitting* given the reasons available to the agent. By "fitting", I mean something along the lines of appropriate, merited, proper, rational, or warranted. For instance, a belief that P may be fitting if the person has

[62] I am not assuming here that Fischer and Ravizza's is the right account of moral responsibility. Several philosophers have made plausible objections to the view, for reasons that are out of the scope of this paper.

[63] There are also some requirements that would need to be dropped. For instance, Fischer and Ravizza emphasize that, for a person to have moral responsibility for an action, the mechanism(s) producing it must not have been tampered with externally. But maybe we do not need this clause for psychiatry, if substance abuse (or being the victim of violent crime) can be seen as external tampering, and it can cause a mechanism to be less reasons-responsive.

encountered sufficient evidence of the truth of P. An emotion, such as fear, may be unfitting if it's produced as a response to something regarded as not dangerous, like a puppy. A desire to Y may be unfitting if the person believes that getting Y would be, all things considered, bad for her. And so on. While "action" is the primary locus of normative ethical analysis, folk psychology issues explanations and predictions not only of action, but *also* of mental states, attitudes, moods, and so on.

- **Reasons**. The relevant reasons in the case of the "mental" won't be restricted to moral ones: there may also be epistemic reasons, practical reasons, instrumental reasons, reputational reasons, and so on. Again, folk-psychology is not restricted in considering only moral reasons for action or for a mental state, so this expansion is pertinent.

The regulative ideal for psychiatry, as a background assumption that is not made explicit but nonetheless guides the approaches of clinicians and researchers, won't be an account with necessary and sufficient conditions for psychopathology. It will probably have exceptions, given the social nature of inclusion of disorders on the DSM. But we can still say some things about what paradigmatic cases of mental illness involve. Consider the mechanism(s) that typically underlie the production of certain mental outputs. A mental disorder is a condition that makes such mechanism(s) extremely less reasons-responsive, compared to what we would otherwise expect it to be. For instance, what separates a person who is afraid of spiders from an arachnophobe are the counterfactual scenarios in which the mechanism(s) that would make them touch a spider would operate. The mechanism(s) behind a person who is merely afraid, compared to those of an arachnophobe, would

be more sensitive and reactive to sufficient reasons to touch the spider. Rational convincing that the spider in question is actually harmless, for example, may have some effect (however small) in the person who is afraid, but not on the one who has a phobia to spiders. Thus, in the case of mental disorders, we expect the individual to consistently display outputs that are unfitting or inappropriate given the reasons available to them. Those unfitting mental states or actions are the symptoms of the disorder in question.

Consider, for example, depression. Depression can be regarded as a persistent negative mood despite what in other people would be sufficient reasons for the contrary one. Phobias are marked fears or anxieties surrounding a certain object or situation, that cannot be modulated by rational arguments. These responses persist even after recognizing they are unwarranted. People with PTSD avoid situations that remind them of the traumatic event experienced, even if they know there is nothing dangerous in those new situations. Delusions involve beliefs that robustly persist despite sufficient evidence to the contrary (Flores, 2021). The loss of control in addiction involves a loss of responsiveness to sufficient reasons to not consume (Burdman, 2022). Sexual dysfunctions involve difficulty getting aroused that are not attributable to there being sufficient reasons for it, such as relationship distress. People with pica cannot help but eat non-food substances despite being aware that such behavior is detrimental to their health. Histrionic personality disorder involves an excess of emotionality that is not warranted by the events. In all these cases, what seems common is that the mechanisms typically producing such emotions, beliefs, or behaviors display a lack of reasons-responsiveness.

Note that, since my account of mental disorders as extreme forms of lack of reasons-responsiveness is based on a folk psychological understanding of the mind, it can be a correct characterization of psychiatric practice even if it tuns out that the production of certain mental states and behaviors is, in fact, not reasons-responsive. That is, my account of mental disorders doesn't make metaphysical assumptions of what is in fact involved behind the symptoms of psychopathologies —only of what is involved when thinking of certain symptoms as psychiatric disorders.

4. Considerations favoring my account of mental disorders

In the last section, I laid down the regulative ideal that, in my opinion, guides psychiatry's inclusion (and exclusion) of certain clusters of symptoms as mental disorders. In this section, I provide some considerations supporting my claim. The best evidence that the account presented here correctly depicts the regulative ideal behind psychiatry is that it explains certain features of the DSM that were previously unaccounted for. The most plausible case for my view is that it makes sense of several previously puzzling aspects of the DSM.

First of all, why do a majority of DSM diagnoses require a person to consistently display the symptoms of a disorder for a considerable amount of time? This requirement for diagnosis does not have a counterpart in medicine: it would be ridiculous for a doctor to ask a newly struggling patient to come back after a few weeks in order to diagnose whether the cause of her novel symptoms is a fracture, a

bacterial infection or a cancerous tumor. Yet, psychiatric diagnoses often require that the patient has been experiencing the symptoms for a considerable amount of time. For instance, an individual must have been consistently exhibiting symptoms for at least two weeks to have Major Depressive disorder, a year to have Disruptive Mood Dysregulation disorder, 3 months for Insomnia, and a month for Substance-Induced Obsessive-Compulsive disorder, and so on. Such requirements don't make much sense on a "realist" view of mental disorders, especially if one considers psychopathologies to correspond to biological occurrences; what is the wait for? They also seem arbitrary in more constructivist or pragmatist accounts, especially since it would be more practical to treat symptoms right away. But these historical requirements make perfect sense under my view. As I said, psychopathologies occur when the degree of reasons-responsiveness of the mechanism(s) responsible for the production of certain mental outputs is significantly diminished —at least, with respect to what we would folk-psychologically expect it to be. To capture the degree to which a mechanism is reasons-responsive, we would have to observe its behavior under different scenarios: does the mechanism produce different outputs at least some of the times there are sufficient reasons to do so? Obviously, an answer in either direction presumes that we have access to the mechanism's outputs in a range of situations. Given that clinicians are often not in a position to intervene in the daily scenarios that a person confronts, they have to use something else as an indicator of the lack of reasons-responsiveness of a mechanism: a robust pattern of unfitting mental outputs over an extended period of time. In other words: since the reasons-responsiveness of a mechanism is not directly epistemically accessible to

psychiatrists, they have to use an indirect indicator of it. Extended periods of time would normally be correlated with a higher number of different scenarios, so if one's symptoms persist over a significant period of time, it is likely that the cause is the mechanism itself, rather than a lack of sufficient reasons to act, feel or believe otherwise.

A second puzzling feature of DSM diagnoses are that they include non-medical exclusionary criteria —which, on my view, track situations that can be construed as the person's having "sufficient reasons" for displaying certain symptoms. To illustrate: a diagnosis of Major Depression disorder required, up until DSM-IV, that one had not suffered the death of a loved one in the two months prior. This "bereavement exclusion clause" was removed from the DSM-5, but the manual still cautions clinicians against mis-diagnosing in such cases. For instance, the DSM remarks "An expectable or culturally approved response to a common stressor or loss, such as the death of a loved one, is not a mental disorder" (DSM-5, Introduction), and it urges clinicians to consider a patient's history and cultural practices of grieving when considering whether a patient has depression in addition to the normal responses in those circumstances[64]. Exclusionary clauses are not just found for depressive disorders: most DSM-5 diagnoses include exclusionary criteria in the form

[64] The Diagnostic criteria section for Major Depressive Disorder includes the following note: "Responses to a significant loss (e.g., bereavement, financial ruin, losses from a natural disaster, a serious medical illness or disability) may include the feelings of intense sadness, rumination about the loss, insomnia, poor appetite, and weight loss noted in Criterion A, which may resemble a depressive episode. Although such symptoms may be understandable or considered appropriate to the loss, the presence of a major depressive episode in addition to the normal response to a significant loss should also be carefully considered. This decision inevitably requires the exercise of clinical judgment based on the individual's history and the cultural norms for the expression of distress in the context of loss" (DSM-5).

of culture-specific reasons for the mental state or action in question (e.g. the value placed to strong interdependence among family members in certain cultures would preclude some such individuals from a diagnosis of Separation Anxiety disorder, even if significant distress or impairment results from such separation). Clinicians may also refuse to diagnose a patient in distress if they judge the patient has a rational basis for her mental state or action (e.g. a person in an abusive relationship may not be depressed, despite showing all the symptoms[65]). Yet, exclusionary clauses remain controversial, since in practice they exclude people who could benefit from psychiatric treatment from receiving it. So, what is the basis for such non-medical exclusionary criteria? My account provides a rationale for them: if a mental state or action is a fitting response to the circumstances, then it does not signal lack of reasons-responsiveness of the producing mechanism, despite being impairing or involving distress. Thus, non-medical exclusionary criteria play the role of separating reasons-responsive mental states and behavior from those that are not.

Third, my account explains the initial plausibility of the idea that mental illness at least mitigates moral responsibility. There is a debate as to whether people with mental disorders are morally responsible for their actions when those are affected by the psychopathology in question, both in ethics (see, e.g. Pickard, 2011; King & May, 2018; Kozuch & McKenna, 2016) and in law (e.g. Elliott, 1996; Kalis & Meynen, 2014). If moral responsibility requires guidance control, but the mechanism(s)

[65] For another hypothetical example, a person who is distressed and removes herself from social situations after knowing she is in a hit list from ISIS may come to believe that she is being followed and be constantly fearful of that. Those beliefs and fears, however, are appropriate given her circumstances —even if they turn out to be false—, differently from those of someone with persecutory delusions.

producing action A in a person with a psychopathology are responsive to reasons to a significant lesser degree, then it is plausible that her degree of moral responsibility for action A is correspondingly diminished. That is, if a mental disorder affects certain mental states and actions, it seems plausible that a person is less blameworthy for the affected states and actions. The reason these debates exist is because both the notion of moral responsibility and that of mental disorders have in the background, for many, the common notion of reasons-responsive mechanisms. This said, my account does not entail that a person with a mental disorder is never responsible for her actions impacted by the psychopathology. A person may retain certain degree of guidance control that would allow her to not have done that particular action at that time. Moreover, even if a mechanism were to be not reasons-responsive at all, it may still be possible for her to have acted otherwise. If a person has access to treatment that allows her to modulate the workings of the affected mechanism(s) (e.g. by taking medication, or doing cognitive behavioral therapy), we may hold her morally responsible for failing to have treatment and therefore, for the actions that she engaged in while disordered. This won't be dissimilar to how we hold people acting under the influence of alcohol as responsible for their actions. Despite knowing that being drunk momentarily impairs one's psychological mechanisms (e.g. decision making, reflexes, and so on), we still hold drunk drivers responsible for the accidents they caused (maybe more so!), since they should have foreseen that drinking would affect their driving skills yet they still chose to drive under influence. In any case, a discussion of the implications of psychopathology for moral responsibility is outside of the scope of this paper.

Finally, my account also makes sense as to why certain pathologies of perception are not included in the DSM, despite affecting states we would folk-psychologically classify as "mental". Folk psychology regards perceptual states as the mere *causal* products of certain inputs, in a way that is not subject to guidance control. Some philosophers even regard as matter of conceptual truth that, if a subject perceives X, then necessarily X is the cause of the subject's perceptual state (e.g. Grice, 1961). Regardless of the merits of such claim, it serves to illustrate that perceptual states, differently from other mental states, are not generally thought of as reasons-responsive. It should then be unsurprising that pathologies affecting especially perception, such as prosopagnosia, phantom limb syndrome, or ear ringing are not consider *mental* disorders, but pathologies of another sort. This is the case even if they are grounded in neurological disturbances and they result in abnormal phenomenology. The account provided explains such exclusions. Nonetheless, one may point out, perceptual states influence subsequent action and mental states, so they have a role in the workings of reasons-responsive mechanisms. And what sets pathologies of perception apart from psychopathologies involving hallucinations and delusions is that such subsequent mechanisms are still normally reasons-responsive in the former case, but not for the second.

So far, I offered four reasons in favor of my account, in the form of phenomena related to DSM diagnostic categories that were left unexplained by other theories. If we understand mental disorders as conditions diminishing the reasons-responsiveness of certain mechanisms producing mental outputs, we can make sense of (i) requirements concerning the period of time a person must have been experiencing the

symptoms; (ii) non-medical exclusionary criteria; (iii) the intuitive link between psychopathologies and a reduced degree of moral responsibility; and (iv) why disorders of mere perception are not included in the DSM.

5. Why does psychiatry restrict itself to disorders of reasons-responsiveness?

If my account of mental disorders as extreme cases of lack of reasons-responsiveness is correct, it raises a puzzle. Why has psychiatry adopted reasons-responsive mechanisms as the locus for psychopathologies, as opposed to e.g. any *mental* mechanism, including perceptual ones? I think the answer is historical, having to do with the historical factors surrounding the development of psychiatry as its own sub-discipline within medicine.

In the early twentieth century, Sigmund Freud transformed the understanding of mental disorders —which until then wasn't uniform across Western Europe[66]. Freud considered that deviant mental states and behaviors that were distressing for the person experiencing them (such as worries, anxieties and obsessions), were the

[66] For instance, during the XIX century in Spain there were two schools of psychiatric thought: an idealist one, which considered that alterations of the soul were involved in mental illness and its treatment; and a materialist one, inspired by a similar German school, which considered mental illness the product of brain deformities. Both, however, favored a "moral treatment" to their patients, involving a favorable environment, often organized by monks and involving faith (Le Bow, 1964). These schools were quite different from what was going on in the UK, where considerations of insanity seemed to go hand-in-hand with involuntarily admissions to asylums —which included not only these with e.g. personality disorders or cognitive disabilities, but also alcoholics, paupers, and promiscuous pregnant single women (Rollin, 2003), which later eugenics would come to see not as mentally "ill", but reflecting the personal trait of having low intelligence. Differently from the UK, in Italy, asylums only admitted lunatics in need of "hard" custody from the 1820s, that were to be seen by specialized medical doctors (De Rissio, 2019). Freudian psychoanalysis unified psychiatric practice across countries.

product of hidden conflicts —and to be treated using a psychoanalysis (Graham, 2010). Psychoanalysis is a sort of verbal therapy consisting into delving into a person's past experiences and thoughts to come upon unconscious feelings, beliefs and memories that may influence one's current deviant mental states and behavior — and then "tackle" them with the help of the therapist. Freud's approach contains some of the ingredients of the account of mental disorders sketched in this paper: we can understand the mental states and behaviors of interest for psychoanalysis as those that are produced by a mechanism which got "stuck" in their response pattern due to some "reason" in the past, like childhood traumatic experiences that became "repressed". The underlying assumption seems to be that normally those mental states and behaviors are produced in response to proximal reasons, and that the hope for treatment is to re-assess the underlying "reasons" behind the production of such mental states via conscious reasoning with the help of the therapist. The Freudian approach dominated psychiatric theory and practice until the 1960s, when it became heavily criticized as psychiatry took a more scientific and neuro-biological turn. Despite such changes, it seems that the subject matter of psychiatry was already inherited from Freud: the target were those deviant mental states and behaviors that were usually produced by reasons-responsive psychological processes, and someone's having a mental illness was not restricted to severe cases as before, but also included more mild ones where the person was experiencing significant distress or disability.

A second possible factor behind psychiatry's regulative-ideal of psychopathologies could come from background assumptions of what separates humans and non-human

animals ("animals", for short). Since René Descartes, animals have been regarded as reflex-driven, machine-like "automata" lacking reason. To act on the basis of reasons, for Descartes, meant to be able to apply general principles to an open-ended set of circumstances. Differently from humans, animals were thought to be incapable of acting *for* reasons, since they showed no evidence of being able to transfer general principles-based knowledge to novel circumstances (Lurz, 2022). But their perceptions, like those of humans, could be causally explained. Even as our understanding of animal minds developed, considerations like Morgan's canon[67] have precluded the use of reasons-involving language to describe their psychological states and processes, favoring associationist and merely causal explanations instead. Behaviorism has maintained as the dominant approach for the study animal minds, even as human psychology took a cognitive turn. Our understanding of animal minds consists mainly in that of the relations between their inputs and outputs, perception and behavior. In contrast, appealing to reasons for mental states and actions became scientifically legitimate in the human case. Maybe psychiatry has been influenced by such context during its development in the XXth century, paying increasingly more attention to reasons-responsive occurrences, regarding the difference between our minds as not being of degree, but of kind. Yet, human and animal minds share perception and action. This might have influenced psychiatry against including disorders of perception (e.g. congenital blindness, prosopagnosia, tinnitus) among its subject matter. But everyone allows that animals act, so if psychiatry is restricted to

[67] The classical formulation of Morgan's canon is: "In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale" (Morgan 1894, p. 53)

disorders of the human mind, why are disorders of behavior (e.g. Parkinson's, kleptomania, Tic disorders) are in the DSM? The answer is, I think, because human actions are characteristically reasons-responsive —or, at least, we folk-psychologically take them to be.

The importance and influence of these factors is, I contend, somewhat speculative. In any case, the main goal for this paper wasn't to examine the historical development of the discipline of psychiatry, but rather to provide a plausible hypothesis of what mental disorders generally are. I hope the account provided here helps elucidate some of the outstanding puzzles in philosophy of psychiatry.

6. Conclusion

What are mental disorders? In this paper, I argued that to understand mental illness we ought to look at the regulative ideal behind the practice of psychiatry. This regulative ideal relies on folk psychology to delineate which alterations of functioning fall within its domain. According to folk psychology, some mental states (such as beliefs, emotions, moods, worries) and behaviors are produced by psychological mechanisms that usually are *reasons*-responsive. This implies that, in some counterfactual scenarios where I had sufficient reasons to act otherwise (or to have a different mental state), I would have done so. A mental disorder is an alteration of the degree of reasons-responsiveness of such mechanism(s), such that

they become significantly insensitive to the relevant reasons, significantly non-reactive to them, or both.

This account of mental disorders makes sense of some features of DSM diagnoses that were previously insufficiently accounted for, such as symptom-exhibition time requirements, exclusionary criteria that don't have a medical basis, the link between psychopathologies and a reduced degree of moral responsibility, and why mere pathologies of perception are not included in the DSM, whereas pathologies of action are.

Bibliography

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical psychology*, *14*(1), 43-64.

Adams, F., & Aizawa, K. (2012). Defending the bounds of cognition. In *The Extended Mind* (pp. 67-80). Boston, MA: MIT Press.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Anscombe, G. E. M. (1971). "Causality and Determination," reprinted in E. Sosa and M. Tooley (eds.), *Causation*. Oxford: Oxford University Press, 1993, pp. 88-104.

Atkinson, A. P., & Adolphs, R. (2011). The neuropsychology of face perception: beyond simple dissociations and functional selectivity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1726-1738.

Atran, S. (2001). A cheater-detection module? Dubious interpretations of the Wason selection task and logic. *Evolution and Cognition*, *7*(2), 187-192.

Baron-Cohen S. (1995) *Mindblindness: an essay on autism and theory of mind.* Boston: MIT Press/Bradford Books.

Barrett, H. C. (2014). *The shape of thought: How mental adaptations evolve*. Oxford University Press.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological review*, *113*(3), 628.

Bate, S., Haslam, C., Tree, J. J., & Hodgson, T. L. (2008). Evidence of an eye movement-based memory effect in congenital prosopagnosia. *Cortex*, *44*(7), 806-819.

Bechtel, W. (2008a). Mechanisms in cognitive psychology: What are the operations?. *Philosophy of Science*, *75*(5), 983-994.

Bechtel, W. (2008b). Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience. London, UK: Routledge

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36(2), 421-441.

Bechtel, W., & Richardson, R. C. (1993). Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research. Princeton: Princeton University Press.

Bechtel, W., & Shagrir, O. (2015). The non-redundant contributions of Marr's three levels of analysis for explaining information-processing mechanisms. *Topics in Cognitive Science*, *7*(2), 312-322.

Bermudez, J. L. (2004). *Philosophy of Psychology: A Contemporary Introduction*. Routledge.

Berridge, K. C., & Kringelbach, M. L. (2013). Neuroscience of affect: brain mechanisms of pleasure and displeasure. *Current opinion in neurobiology*, *23*(3), 294-303.

Block, N. J., & Fodor, J. A. (1972). What psychological states are not. *The Philosophical Review*, *81*(2), 159-181.

Bolton, D., & Hill, J. (1996). *Mind, meaning, and mental disorder: The nature of causal explanation in psychology and psychiatry.* Oxford University Press.

Borsboom, D. (2017), A network theory of mental disorders. *World Psychiatry*, 16: 5-13.

Brentano, F. (1874). *Psychology From an Empirical Standpoint*. Routledge.

Burdman, F. (2022). A pluralistic account of degrees of control in addiction. *Philosophical Studies*, *179*(1).

Calvo, P., & Gomila, A. (Eds.). (2008). *Handbook of cognitive science: An embodied approach*. Oxford, UK: Elsevier.

Camp, E. (2007). Thinking with maps. *Philosophical Perspectives* 21 (1):145–182.

Carruthers, P. (2006). *The architecture of the mind*. Oxford University Press.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press.

Centanni, T. M., Norton, E. S., Park, A., Beach, S. D., Halverson, K., Ozernov-Palchik, O., & Gabrieli, J. D. (2018). Early development of letter specialization in left fusiform is associated with better word reading and smaller fusiform face area. Developmental Science, 21(5), e12658.Barrett, B., Muller, D., Rakel, D., Rabago, D., Marchand, L., & Scheder, J. C. (2006). Placebo, meaning, and health. *Perspectives in Biology and Medicine*, *49*(2), 178-198.

Charland, L. C. (2004). Personality Disorders. *The philosophy of psychiatry: A companion*, 64.

Charland, L. C. (2006). Moral nature of the DSM-IV Cluster B personality disorders. *Journal of personality disorders*, *20*(2), 116-125.

Chawner, S., & Owen, M. J. (2022). Autism: A model of neurodevelopmental diversity informed by genomics. *Frontiers in psychiatry*, *13*, 981691. https://doi.org/10.3389/fpsyt.2022.981691

Chemero, A., & Silberstein, M. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of science*, *75*(1), 1-27.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58*(1).

Clark, A., & Toribio, J. (1994). Doing without representing?. *Synthese*, *101*(3), 401-431.

Colloca, L., & Miller, F. G. (2011). How placebo responses are formed: a learning perspective. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *366*(1572), 1859–1869.

Coltheart, M. (1999). Modularity and cognition. *Trends in cognitive sciences*, *3*(3), 115-120.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. *The adapted mind: Evolutionary psychology and the generation of culture*, *163*, 163-228

Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science*, 68(1):53–74.

Craver, C. F. (2006). When mechanistic models explain. *Synthese,* 153 (3):355-376.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford and New York: Oxford University Press.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, 22(5):575–594.
Craver, C. F. (2014). The ontic account of scientific explanation. In Kaiser, M. I., Scholz, O. R., Plenge, D., and Hüttemann, A., editors, *Explanation in the Special Sciences: The Case of Biology and History*, pages 27–52. Springer Verlag.
Craver, C. F. & Darden, L. (2001). Discovering mechanisms in neurobiology: The case of spatial memory. In P.K. Machamer, Rick Grush & Peter McLaughlin (eds.), *Theory and Method in Neuroscience*. Pittsburgh: University of Pitt Press. pp. 112--137.
Craver, C. F. and Kaplan, D. M. (2011). Towards a mechanistic philosophy of neuroscience. In French, S. and Saatsi, J., editors, *Continuum Companion to the Philosophy of Science*, page 268. London: Continuum.
Craver, C. F. and Wilson, R. A. (2006). Realization. In Thagard, P., editor, *Handbook of the Philosophy of Psychology and Cognitive Science*. Elsevier.
Craver, C. F., & Tabery, J. (2019). "Mechanisms in Science", The Stanford Encyclopedia of Philosophy (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/>.
Craver, C. F., Glennan, S. & Povich, M. (2021). Constitutive relevance & mutual manipulability revisited. *Synthese* 199 (3-4):8807-8828.
Craver, C. F. & Darden, L. (2001). Discovering mechanisms in neurobiology: The case of spatial memory. In P.K. Machamer, Rick Grush & Peter McLaughlin (eds.), *Theory and Method in Neuroscience*. Pittsburgh: University of Pitt Press. pp. 112--137.
Cummins, R. (1983). *The Nature of Psychological Explanation*. Boston, MA: MIT Press.
Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine*, *11*(1), 1-8.
Darden, L. (1996). Generalizations in biology. *Studies in History and Philosophy of Science Part A*, *27*(3).
Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science*, *69*(S3), S354-S365.
Darden, L. (2006). *Reasoning in Biological Discoveries: Essays on Mechanisms, Interfield Relations, and Anomaly Resolution*. Cambridge University Press.
Darden, L. (2008). Thinking Again about Biological Mechanisms. *Philosophy of Science, 75*(5), 958-969. doi:10.1086/594538
Darden, L., & Craver, C. F. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *33*(1), 1-28.
Darden, L. and Craver, C. F. (2013). *In Search of Biological Mechanisms: Discoveries across the Life Sciences. Chicago*. IL: University of Chicago Press.

Darden, L., Pal, L. R., Kundu, K., & Moult, J. (2018). The product guides the process: discovering disease mechanisms. In Danks, D. & E. Ippoliti (eds.), *Building theories* (pp. 101-117). Springer International Publishing..

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy*, *60*(23), 685–700. https://doi.org/10.2307/2023177

Davidson, D. (1970). Mental Events. Reprinted in D. Davidson (1980). *Essays on Actions and Events*, 201-224.

Dawson, M. R. (1998). *Understanding cognitive science*. Oxford, UK: Blackwell Publishing.

Dawson, M. R. (2013). *Mind, body, world: foundations of cognitive science*. Edmonton, AB: Athabasca University Press.

de la Fuente-Fernández, R. (2009). The placebo-reward hypothesis: dopamine and the placebo effect. *Parkinsonism & Related Disorders*, *15*, S72-S74.

De Rissio, A (2019). *The Italian Psychiatric Experience*. Blackwell UK.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Filho, G. N., Jobert, A., & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science*, *330*(6009), 1359-1364.

Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy* 68 (February):87-106.

Dennett, D. C. (1985). *Brainstorms*, Cambridge, MA: MIT Press.

Dennett, D. C. (1987). *The Intentional Stance*, Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness Explained*. Penguin Books.

Dennett, D. C. (1992). The self as a center of narrative gravity. In Frank S. Kessel, P. M. Cole & D. L. Johnson (eds.), *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum. pp. 4-237.

Destexhe, A. (2012). *Neuronal noise*. New York: Springer.

Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives*, 26.

Drayson, Z. (2015). The Philosophy of Phenomenal Consciousness. In *The Constitution of Phenomenal Consciousness*. Amsterdam: pp. 273-292.

Dretske, F. (2004). Psychological vs. biological explanations of behavior. *Behavior and Philosophy* 32 (1):167-177.

Eichenbaum, H., Yonelinas, A. R., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual review of neuroscience*, *30*, 123.

Elliott, C. (1996). *The rules of insanity: Moral responsibility and the mentally ill*. SUNY Press.

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience*, *20*(11), 1504-1513.

Eysenck, M. W., & Keane, M. T. (2015). *Cognitive psychology: A student handbook*. Psychology Press.

Falguera, J. L., Martínez-Vidal, C., & Rosen, G. (2022) “Abstract Objects", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2022/entries/abstract-objects/>.

Fine, A. (2001) The Scientific Image Twenty Years Later. *Philosophical Studies* 106, 107–122.

Fischer, J. M. (1994). *The Metaphysics of Free Will: An Essay on Control*. Wiley-Blackwell.
Flores, C. (2021). Delusional evidence-responsiveness. *Synthese* 199 (3-4):6299-6330.
Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2):97–115.
Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT press.
Fodor, J. A., & McLaughlin, B. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35(2), 183–204.
Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3-71.
Fodor, J., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*(2), 183-204.
Foucault, M. (1975). *The Birth of the Clinic: An Archaeology of Medical Perception.* New York: Vintage Books.
Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy*, *66*(23), 829–839. https://doi.org/10.2307/2023833
Franklin-Hall, L. R. (2016). New Mechanistic Explanation and the Need for Explanatory Constraints. In K. Aizawa & . Gillett (eds.), *Scientific Composition and Metaphysical Ground*, UK: Palgrave Macmillan. pp. 41-74.
Frisaldi, E., Shaibani, A., & Benedetti, F. (2020). Understanding the mechanisms of placebo and nocebo effects. *Swiss Medical Weekly*, (35).
Galaj, E., & Ranaldi, R. (2021). Neurobiology of reward-related learning. *Neuroscience and biobehavioral reviews*, *124*, 224–234.
Gall, F. J. (1835). On the functions of the brain and of each of its parts: With observations on the possibility of determining the instincts, propensities, and talents, or the moral and intellectual dispositions of men and animals, by the configuration of the brain and head (Vol. 1). Marsh, Capen & Lyon.
Garson, J. (2017). A generalized selected effects theory of function. *Philosophy of Science*, *84*(3), 523-543.
Garson, J. (2022a). Madness: A Philosophical Exploration. Oxford University Press.
Garson, J. (2022b). *Seeing dyslexia as a unique cognitive strength, rather than a disorder*. Psychology Today. Retrieved September 21, 2022, from https://www.psychologytoday.com/us/blog/the-biology-human-nature/202207/seeing-dyslexia-unique-cognitive-strength-rather-disorder
Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision research*, *37*(12), 1673-1682
Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. Nature neuroscience, 6(4), 428-432.
Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1), 49-71.
Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of science*, *69*(S3), S342-S353.

Glennan, S. (2010). Ephemeral mechanisms and historical explanation. *Erkenntnis*, *72*(2), 251-266.
Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.
Glennan, S., & Illari, P. (2017). Varieties of mechanisms. In *The Routledge handbook of mechanisms and mechanical philosophy*(pp. 91-103). London, UK: Routledge.
Glennan, S., Illari, P., & Weber, E. (2022). Six Theses on Mechanisms and Mechanistic Science. *Journal for General Philosophy of Science*, *53*(2), 143-161.
Godfrey-Smith, P. (2009). Causal pluralism. In Beebee, H., Menzies, P., and Hitchcock, C., editors, *The Oxford Handbook of Causation*, pages 326–337. Oxford University Press.
Goldstein, A. N., & Walker, M. P. (2014). The role of sleep in emotional brain function. *Annual review of clinical psychology*, *10*, 679.
Grabenhorst, F., & Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in cognitive sciences*, *15*(2), 56-67.
Graham, G. (2010). *The Disordered Mind: An Introduction to Philosophy of Mind and Mental Illness*. Routledge.
Grice, H. P. (1961). The causal theory of perception. In Jonathan Dancy (ed.), *Aristotelian Society Supplementary Volume*. Oxford University Press. pp. 121-168.
Griffiths, P. E. (1993). Functional analysis and proper functions. *The British Journal for the Philosophy of Science*, *44*(3), 409-422.
Hacking, I. (1999). *The Social Construction of What?*. Harvard University Press.
Haji, I. (1998). *Moral Appraisability*, New York: Oxford University Press.
He, S., Liu, H., Jiang, Y., Chen, C., Gong, Q., & Weng, X. (2009). Transforming a left lateral fusiform region into VWFA through training in illiterate adults. *Journal of Vision*, *9*(8), 853-853.
Healy, S. E. (1998). *Spatial representation in animals.* Oxford University Press.
Herrnstein, R. J. (1962). Placebo effect in the rat. *Science*, *138*(3541), 677-678.
Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*, Cambridge, MA: Harvard University Press.
Hochstein, E. (2019). How metaphysical commitments shape the study of psychological mechanisms. *Theory & Psychology*, *29*(5), 579-600.
Hole, G., & Bourne, V. (2010). *Face processing: Psychological, neuropsychological, and applied perspectives*. Oxford University Press.
Illari, P. M., & Williamson, J. (2010). Function and organization: Comparing the mechanisms of protein synthesis and natural selection. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *41*(3), 279-291.
Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119-135.
Kaiser, M. I. (2017). The Components and Boundaries of Mechanisms. In S. Glennan & P. Illari (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. New York, USA: Routledge.

Kaiser, M. I. & Krickel, B. (2017). The Metaphysics of Constitutive Mechanistic Phenomena. *British Journal for the Philosophy of Science* 68 (3).
Kalis, A., & Meynen, G. (2014). Mental disorder and legal responsibility: The relevance of stages of decision making. *International journal of law and psychiatry*, *37*(6), 601-608.
Kang, D. W., Adams, J. B., Coleman, D. M., Pollard, E. L., Maldonado, J., McDonough-Means, S., & Krajmalnik-Brown, R. (2019). Long-term benefit of Microbiota Transfer Therapy on autism symptoms and gut microbiota. *Scientific reports*, *9*(1), 1-9.
Kanwisher, N. (2000). Domain specificity in face perception. *Nature neuroscience*, *3*(8), 759-763.
Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of science*, *78*(4), 601-627.
Kendell, R. E. (1991). Relationship between the DSM-IV and the ICD-10.. *Journal of Abnormal Psychology, 100*(3), 297–301.
Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders?. *Psychological medicine*, *41*(6), 1143–1150.
Kim, J. (1988). "Explanatory Realism, Causal Realism, and Explanatory Exclusion". *Midwest Studies in Philosophy*, 12, p. 225-239.
Kim, J. (1989). "Mechanism, Purpose, and Explanatory Exclusion". *Nous-Supplement*: *Philosophical Perspectives*, 3, p. 77-108.
Kim, J. (1998). *Mind in a Physical World.* Cambridge: MIT Press.
King, M., & May, J. (2018). Moral responsibility and mental illness: A call for nuance. *Neuroethics*, *11*(1), 11-22.
Kozuch, B., & McKenna, M. (2015). Free Will, Moral Responsibility, and Mental Illness. In *Philosophy and Psychiatry* (pp. 105-129). Routledge.
Krickel, B. (2018). *The Mechanical World: The Metaphysical Commitments of the New Mechanistic Approach*. Springer Verlag.
Latuske, P., Kornienko, O., Kohler, L., & Allen, K. (2018). Hippocampal Remapping and Its Entorhinal Origin. *Frontiers in behavioral neuroscience*, *11*,
Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, *8*(1), 98-113.
Le Bow, R. H. (1964). Spain And Psychiatry In The Latter Part Of The 19th Century. *Bulletin of the History of Medicine*, *38*(5), 444–454.
Levine, J., Gordon, N., & Fields, H. (1978). The mechanism of placebo analgesia. *The Lancet*, *312*(8091), 654-657.
Levitin, D. J. (2002). *Foundations of Cognitive Psychology: Core Readings*. MIT Press.
Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental science*, *14*(6), 1292-1300.
Lopatina, O. L., Komleva, Y. K., Gorina, Y. V., Higashida, H., & Salmina, A. B. (2018). Neurobiological aspects of face recognition: The role of oxytocin. *Frontiers in behavioral neuroscience*, *12*, 195.

Lurz, R. (2022). Animal minds. The Internet Encyclopedia of Philosophy. Available at: https://iep.utm.edu/animal-mind/. Accessed Oct. 3rd 2022.
Machamer, P. (2004). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science*, 18(1):27–39.
Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25.
Mazzeo, S. E., & Bulik, C. M. (2009). Environmental and genetic risk factors for eating disorders: what the clinician needs to know. *Child and adolescent psychiatric clinics of North America*, *18*(1), 67–82.
Meyer, R. (2018). The nonmechanistic option: Defending dynamical explanation. *British Journal for the Philosophy of Science*.
Meyer, R. (2020). The Non-mechanistic Option: Defending Dynamical Explanations. *British Journal for the Philosophy of Science* 71 (3):959-985.
Michaelian, K. (2011). Is memory a natural kind? *Memory Studies* 4 (2):170-189.
Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Boston, MA: MIT Press.
Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science* 56 (June):288-302.
Millikan, R. G. (2002). Biofunctions: Two paradigms. In Andre Ariew (ed.), *Functions*. Oxford University Press. pp. 113-143.
Montgomery, G. H., & Kirsch, I. (1997). Classical conditioning and the placebo effect. *Pain*, *72*(1-2), 107–113.
Morgan, C. L. (1894). *An Introduction to Comparative Psychology*. New York: Scribner.
Moring, B. (2014). *Research methods in psychology: Evaluating a world of information*. WW: Norton & Company.
Murphy, D. (2006). *Psychiatry in the scientific image*. MIT Press.
Nagel, T. (1974). What is it like to be a bat?. *The philosophical review*, *83*(4), 435-450.
Nelkin D. (2011). *Making Sense of Free Will and Responsibility*, New York: Oxford University Press.
Nicholson, D. J. (2012). The concept of mechanism in biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1):152–163.
Orilia, F. & Paoletti, M.P. (2022). "Properties", *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/properties/>.
Patel, G. H., Kaplan, D. M., & Snyder, L. H. (2014). Topographic organization in the brain: searching for general principles. *Trends in cognitive sciences*, *18*(7), 351–363.
Piccinini, G. and Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183(3):283–311.
Pickard, H. (2009). Mental illness is indeed a myth. In Matthew Broome & Lisa Bortolotti (eds.), *Psychiatry as Cognitive Neuroscience*. Oxford University Press.

Pickard, H. (2017). Responsibility without Blame for Addiction. *Neuroethics*, 10 (1):169-180.
Pinker, S. (1997). *How the mind works* (Vol. 524). New York, NY: Norton.
Plassmann, H., O'Doherty, J., Shiv, B., & Rangel, A. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(3), 1050–1054.
Psillos, S. (2004). A glimpse of the secret connexion: Harmonizing mechanisms with counterfactuals. *Perspectives on science*, *12*(3), 288-319.
Putnam, H. (1967). Psychological predicates. In W.H. Capitan and D.D. Merrill (eds.), *Art, mind, and religion*, *1*, Pittsburgh: University of Pittsburgh Press, 37–48.
Putnam, H. (1975). *Mind, Language and Reality: Philosophical Papers.* Cambridge, UK: Cambridge University Press.
Radden, J. (2019). "Mental Disorder (Illness)", *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2019/entries/mental-disorder/>.
Reddan, M. C., & Wager, T. D. (2018). Modeling pain using fMRI: from regions to biomarkers. *Neuroscience bulletin*, *34*(1), 208-215.
Rey, G. (1997). *Contemporary Philosophy of Mind: A Contentiously Classical Approach*. Wiley-Blackwell.
Rollin, H. (2003). Psychiatry in Britain one hundred years ago. *British Journal of Psychiatry, 183*(4), 292-298. doi:10.1192/bjp.183.4.292
Rupert, R. D. (2009). *Cognitive systems and the extended mind*. Oxford, UK: Oxford University Press.
Samuelson, L. K., Jenkins, G. W., & Spencer, J. P. (2015). Grounding cognitive-level processes in behavior: the view from dynamic systems theory. *Topics in cognitive science*, *7*(2), 191–205.
Sanders, A. E., Slade, G. D., Fillingim, R. B., Ohrbach, R., Arbes, S. J., Jr, & Tchivileva, I. E. (2020). Effect of Treatment Expectation on Placebo Response and Analgesic Efficacy: A Secondary Aim in a Randomized Clinical Trial. *JAMA network open*, *3*(4), e202907.
Sartorio, C. (2016). *Causation and Free Will*, New York: Oxford University Press.
Satcher, D. (2001). *Mental health: Culture, race, and ethnicity—A supplement to mental health: A report of the surgeon general.* US Department of Health and Human Services.
Shagrir, O., & Bechtel, W. (2017). Marr's computational level and delineating phenomena. *Explanation and integration in mind and brain science*, 190-214.
Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.
Shanahan, M. (2016). "The Frame Problem", *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/frame-problem/>.
Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(4), 623–656.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.
Shapiro, L., & Sober, E. (2007). "Epiphenomenalism: The Dos and the Don'ts", in G. Wolters and P. Machamer (eds.), *Studies in Causality: Historical and Contemporary*, Pittsburgh: University of Pittsburgh Press, pp. 235-64.
Shea, N. (2018). *Representation in Cognitive Science*. Oxford, UK: OUP.
Silberstein, M. and Chemero, A. (2012). Complexity and extended phenomenological-cognitive systems. *Topics in Cognitive Science*, 4(1):35– 50.
Silberstein, M. and Chemero, A. (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philosophy of Science*, 80(5):958–970.
Simon, E. B., Rossi, A., Harvey, A. G., & Walker, M. P. (2020). Overanxious and underslept. *Nature Human Behaviour*, *4*(1), 100-110.
Skipper, R. A., & Millstein, R. L. (2005). Thinking about evolutionary mechanisms: natural selection. *Studies in History and Philosophy of Biol & Biomed Sci*, *2*(36), 327-347.
Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes, 102*(2), 143–153.
Smortchkova, J. & Murez, M. (2020). "Representational Kinds". In Smortchkova, J., Dolega, K. & T. Schlicht (eds.), *What are Mental Representations?*. Oxford University Press.
Suchecki, D., Tiba, P. A., & Machado, R. B. (2012). REM sleep rebound as an adaptive response to stressful situations. *Frontiers in neurology*, *3*, 41.
Szasz, T. S. (1960). The myth of mental illness. *American Psychologist, 15*(2), 113–118.
Tabb, K. (2019). Philosophy of psychiatry after diagnostic kinds. *Synthese* 196 (6):2177-2195.
Tolman, E. C., Ritchie, B. F., & Kalish, D. (1946). Studies in spatial learning. I. Orientation and the short-cut. *Journal of experimental psychology*, *36*(1), 13.
Tsou J. Y. (2013). Depression and suicide are natural kinds: implications for physician-assisted suicide. *International journal of law and psychiatry*, *36*(5-6), 461–470. https://doi.org/10.1016/j.ijlp.2013.06.013
Tsou, J. Y. (2022). Biological Essentialism, Projectable Human Kinds, and Psychiatric Classification. *Philosophy of Science*, 1-21.
Tuttle, A. H., Tohyama, S., Ramsay, T., Kimmelman, J., Schweinhardt, P., Bennett, G. J., & Mogil, J. S. (2015). Increasing placebo responses over time in U.S. clinical trials of neuropathic pain. *Pain*, *156*(12), 2616–2626.
Valenza, E., Simion, F., Cassia, V. M., & Umiltà, C. (1996). Face preference at birth. *Journal of experimental psychology: Human Perception and Performance*, *22*(4), 892.
Verdejo, V. M. (2015). The systematicity challenge to anti-representational dynamicism. *Synthese*, *192*(3), 701–722.

Wager, T. D. and Atlas, L. Y. (2015). The neuroscience of placebo effects: connecting context, learning and health. *Nature Reviews Neuroscience*, 16:403–418.
Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C. W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, *368*(15), 1388-1397.
Wakefield, J. C. (1992). Disorder as harmful dysfunction: A conceptual critique of DSM-III-R's definition of mental disorder. *Psychological Review* 99 (2):232-247.
Wakefield, J. C. (1997). Diagnosing DSM-IV—Part I: DSM-IV and the concept of disorder. *Behaviour research and therapy*, *35*(7), 633-649.
Wakefield J. C. (2007). The concept of mental disorder: diagnostic implications of the harmful dysfunction analysis. *World psychiatry: official journal of the World Psychiatric Association (WPA)*, *6*(3), 149–156.
Wallace, R. J. (1996). *Responsibility and the Moral Sentiments,* Cambridge, MA: Harvard University Press.
Webster M. A. (2012). Evolving concepts of sensory adaptation. *F1000 biology reports*, *4*, 21. https://doi.org/10.3410/B4-21
Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese*, 183(3):313–338.
Whitehouse, A. J., & Stanley, F. J. (2013). Is autism one or multiple disorders?. *The Medical journal of Australia*, *198*(6), 302–303.
Wolf, Susan (1990). *Freedom Within Reason*. New York: Oxford University Press.
Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.
Wright, L. (1976). *Teleological explanations: An etiological analysis of goals and functions*. LA, CA: Univ of California Press.
Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, *8*(8), 665-670.
Zachar, P. (2002). The Practical Kinds Model as a Pragmatist Theory of Classification. *Philosophy, Psychiatry, and Psychology* 9 (3):219-227.
Zachar, P. (2014). *A metaphysics of psychopathology*. MIT Press.
Zachar P. (2015). Psychiatric disorders: natural kinds made by the world or practical kinds made by us?. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, *14*(3), 288–290.
Zimmerman, D. (2002). Reasons-Responsiveness and Ownership-of-Agency: Fischer and Ravizza's Historicist Theory of Responsibility. *The Journal of Ethics*, *6*(3), 199–234.