Explanatory integration, computational phenotypes and dimensional psychiatry. The case of alcohol use disorder

Matteo Colombo & Andreas Heinz

Abstract We compare three theoretical frameworks for pursuing explanatory integration in psychiatry: a new dimensional framework grounded in the notion of computational phenotype, a mechanistic framework, and a network of symptoms framework. Considering the phenomenon of alcoholism, we argue that the dimensional framework is the best for effectively integrating computational and mechanistic explanations with phenomenological analyses.

Keywords: alcoholism; computational phenotype; dimensional psychiatry; explanatory integration; mechanism; network of symptoms

1. Introduction

Explanatory integration is one of the traditional aims of psychiatry, but it remains controversial how it should be effectively pursued (Stephan et al. 2016).[1] In this paper, we examine three theoretical frameworks for pursuing explanatory integration in psychiatry: Kendler, Zachar & Craver's (2011) mechanistic framework, Borsboom & Cramer's (2013) network of symptoms framework, and a dimensional framework we are going to develop and ground in the notion of *computational phenotype*, where a computational phenotype is "a measurable behavioural or neural type defined in terms of some computational model" (Montague et al. 2012, 72). Considering alcohol use disorder (AUD) as a case study, we show that in comparison to the mechanistic and network of symptoms frameworks the dimensional framework is more practically useful in a variety of clinical and research contexts, and more adequate for integrating computational and mechanistic explanations with phenomenological analyses.

The motivations for choosing AUD as a case study are threefold. First, AUD is a prominent explanatory target of computational psychiatry. Computational psychiatry aims "to enable integration" of explanations of mental maladies across temporal and spatial scales—from genes to molecules, cells, circuits, brain systems, and individual and social behaviour—"by demonstrating, in a mathematically rigorous way, how phenomena on one level impact phenomena on another" (Kurth-Nelson et al. 2016, 79). Second, AUD has distinctive genetic, neurophysiological, behavioural, social, and cultural correlates, as well as a rich phenomenology. Its phenomenology involves a sense of impaired control over drinking, delirium with delusions and hallucinations, "blackouts", craving, and suffering associated with hangovers, withdrawal and social isolation (see, e.g., Smith 1998; Shinebourne & Smith 2009; Flanagan 2013). Third and finally, given its numerous correlates and its rich phenomenology, and because computational psychiatry has been said to "honour the values and goals of those with lived experience of psychosis" (Powers, Bien, & Corlett 2018, 640), AUD offers an ideal case for assessing how different theoretical frameworks can fruitfully integrate phenomenology, mechanism, and computation in psychiatry.

2. Alcohol use disorder and phenomenology

[1] It is worth clarifying right at the beginning what we take to constitute explanations. All explanations answer some why-, how-, when-, or where-question, although significant variation is observed across scientific and ordinary contexts in what is accepted as an explanation, in what type of explanatory information is sought, and in what norms are assumed to govern good explanations. This apparent variation is reflected in both the philosophy and psychology of explanation (Colombo 2017). For present purposes, we assume that an answer to a request for explanation is a good one to the extent it either unifies apparently scattered pockets of knowledge about the phenomenon of interest, or can be used to address counterfactual questions about what would happen if certain features of the phenomenon of interest were different.

Alcohol is one of the most widely used psychoactive, dependence-producing substances in the world, and is associated with several mental maladies (Connor et al. 2016; WHO 2014). According to the Diagnostic and Statistical Manual of Mental Disorders (*DSM*-5), the criteria for diagnosing alcohol use disorder (AUD) are the following:

1. Alcohol is often taken in larger amounts or over a longer period than was intended.
2. There is a persistent desire or unsuccessful efforts to cut down or control alcohol use.
3. A great deal of time is spent in activities necessary to obtain alcohol, use alcohol, or recover from its effects.
4. Craving, or a strong desire or urge to use alcohol.
5. Recurrent alcohol use resulting in a failure to fulfil major role obligations at work, school, or home.
6. Continued alcohol use despite having persistent or recurrent social or inter-personal problems caused or exacerbated by the effects of alcohol.
7. Important social, occupational, or recreational activities are given up or reduced because of alcohol use.
8. Recurrent alcohol use in situations in which it is physically hazardous.
9. Alcohol use is continued despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to have been caused or exacerbated by alcohol.
10. Tolerance, as defined by either of the following:
    a. A need for markedly increased amounts of alcohol to achieve intoxication or desired effect.
    b. A markedly diminished effect with continued use of the same amount of alcohol.
11. Withdrawal, as manifested by either of the following:
    a. The characteristic withdrawal symptoms for alcohol...
    b. Alcohol [...] is taken to relieve or avoid withdrawal symptoms. (*DSM*-5, 490-1)

Over history, there have been several other diagnostic criteria of alcoholism (Tabakoff & Hoffman 2013). The items on the *DSM*-5 list provide us with a sufficiently typical description of key psychological, behavioural and social aspects of excessive alcohol consumption. Some of these items correlate with various risk factors for AUD, including: aggregate genetic risk factors, impairments in the frontal lobes and their connections with limbic regions in the brain, neuroticism and impulsivity, parental loss, peer alcohol use, and prices of alcoholic beverages (Kendler 2012, 12-14). Given our purposes here, it is important to point out items 4-6 and 9 describe poor learning and decision-making as central features of AUD, and computational modelling of learning and decision-making in AUD has in fact been blooming in the last few years (e.g., Voon, Reiter, Sebold, & Groman 2017); items 2, 4, and 11 indicate AUD patients have expectations, perceptions, desires, moods, and thoughts infused with value, which can exert strong motivational power on their learning and decision-making. Within philosophy, these types of mental states are the targets of phenomenological analyses.

Phenomenology is the study of structures of types of familiar intentional mental states like perceptions, thoughts, emotions, desires, imaginations (i.e., mental states *of*, or *about* something) as they are experienced from a first-person point of view (see Smith 2018 for a comprehensive introduction to phenomenology). One central goal is to develop a holistic account of the lived experience of embodied, ecologically situated agents. Within this broad and heterogeneous field of research, at least three methods have been adopted to study the structures of psychiatric patients' experience. The first method consists in describing a type of lived experience "as it is without taking account of its psychological origin and the causal explanations which the scientist, the historian, or the sociologist may be able to provide" (Merleau-Ponty 1945/1962, vii). Jaspers (1913/1997), for example, provides us with detailed, comprehensive descriptions of patients' lived

experiences, based on biographical information and on notes on how patients felt and thought about their experiences, relationships and condition. Another method adopted by phenomenologists consists in interpreting a type of experience by situating it in a specific social, material and experiential context. For example, Laing (1960) interprets the psychotic experience by situating it in the web of the patients' personal relationships. His interpretation of this type experience is in terms of "ontological insecurity", which is roughly a fragile sense of self impeding people's taking for granted the "realness" and "meaning" of ordinary circumstances of everyday life. A third method consists in analysing the modal structure, and conditions of possibility, of embodied subjective experiences, distinguishing different features of different types of experiences.

Phenomenological methods in psychiatry do not merely consist in interviewing and taking note of patients' reports; they involve complex descriptive, interpretative, and analytical processes, where patients' experiences are organized on the basis of specific theoretical structures (e.g. Carel 2011; Fuchs 2010). For example, Laing (1960) organized patients' experiences on the basis of the existential structure of "ontological insecurity", which, applied to the case of AUD, highlights that alcohol-dependent patients feel more unreal than real, they would feel so separated from the rest of the world they experience their autonomy and identity as constantly threatened, and the only escape from this existential despair is through the anesthetization of alcohol. While triangulating between multiple data sources and methods can contribute to the trustworthiness of phenomenological results, such as descriptions, analyses or interpretations of some mental phenomenon, the ultimate criterion to evaluate these results is their capacity to *make sense* of out-of-the-ordinary, unexpected experiences that cannot be readily understood in terms of more familiar knowledge structures.

Phenomenological descriptions, analyses, or interpretations may not constitute explanations. After all, phenomenology is often characterised as a purely descriptive enterprise distinct from explanation. But this does not mean they are autonomous, that they cannot constrain, inform, or offer as source of evidence for causal, computational, or other types of explanations of mental maladies. In fact, phenomenological results are often used to clarify the structure of the experiences involved in mental maladies, to interpret experimental results, and to inspire hypotheses for further research (Gallagher 2004; Parnas & Sass 2008; Sass 2014).

Consider AUD. One recurrent feature of phenomenological descriptions of AUD is the sense of powerless, helpless suffering that accompanies alcoholics' drinking behaviour. Smith (1998), for example, conducted in-depth interviews with six alcohol-dependent patients, between 42 and 61 years, in a clinic in Scotland. Using interpretative and descriptive phenomenological methods, he puts into focus how patients' "[s]uffering is lived as an insidious process, a movement of ever decreasing circles, whose momentum accelerates you into a rapid, spiralling decline. This vortex is a spinning vicious circle, full of energy, yet symbolising powerlessness" (216).[2]

Smith's descriptions highlight the embodied suffering and experience of self-stigma associated with alcohol-dependent patients' craving and alcohol withdrawal. They also draw our attention to the

[2] Two of Smith's (1998) participants reported: "Suffering is lived in the realization that physical sickness and mental pain increase with each drinking bout and that each bout is an escape from the guilt, shamefulness, and self-loathing set in motion by the previous one. It is lived watching yourself deteriorate in all life aspects, and finding yourself powerless to intervene on your own behalf... You watch helplessly as your addicted self sneaks out to buy more alcohol to finish the job." (216). "Suffering is eventually lived in a state of depression and despair, of powerlessness to break the circle. The eye of the vortex represents the sufferer's own personal rock bottom of physical, social and moral degradation" (Ibid.).

rigidity of behavioural patterns, where patients cannot stop drinking. When they start drinking, every choice they make is often un-flexibly biased towards consuming more alcohol.[3]

Phenomenological descriptions *seem* to suggest that "learning to drink" consists in the acquisition of a rigid habit, of a behavioural pattern that is so ingrained as to make heavy drinkers insensitive to their motivation state and to the bodily, psychological, and social consequences of drinking. In acquiring a habit of drinking, goal-directed processes are likely to be involved too (Everitt & Robbins 2016). But, in acquiring and enacting drinking habits, the flexibility of alcohol-dependent individuals' goal-directed decision processes is often reduced, as alcohol-dependent individuals' repeated drinking behaviour often biases their expectations about the goodness of consuming alcohol. One 31 years old alcohol-dependent patient interviewed by Shinebourne and Smith (2009, 155-6) explains how excessive alcohol consumption reconfigured her sense of self. She says:

"Some big wave, you know, you just get caught with it, that's what it used to be like, this kind of like helpless feeling, just having to go and get drunk almost, you know, not even particularly wanting to, just feeling like there's no other way when you are in that situation. I was very much at sea, really, and, I didn't feel grounded... just this flux and thought, when am I ever going to go on land... and even if you were sitting on the beach, you know, you'd get caught back in..."

This feeling ungrounded, helplessly "caught back in by some big wave," is also one of the recurrent motifs in Flanagan's (2013) memoir of his own alcohol dependence and recovery experience. Relying on his personal experience, but also on literature, Flanagan discusses the myriad ways, in which the habit of drinking blends into one's sense of self. For at least some alcohol-dependent patients, alcohol does not completely hijack their capacity for goal-directed control, but biases it towards drinking. Alcohol becomes part of the way they are: "Their personhood, their character, is constituted, in part, by a history of drinking, by a set of identifications and practices that involve alcohol, and that make these individuals who and what they are. Alcoholism, of this sort, at any rate, is a wide ecological phenomenon; it involves the deep-self" (885-6). Because drinking may be partly constitutive of one's sense of self, "undoing alcoholism as a form of life, and not more narrowly as just a drinking problem, involves fairly radical undoing and then redoing of oneself" (886). It involves acquiring new habits that fill with meaning one's understanding of own subjective experiences and social situation.

In summary, standard diagnostic criteria of AUD indicate that bad decision-making and impaired learning are central features of AUD, and—as we shall see in a moment—computational models of learning and decision-making have been advancing our understanding of AUD. Some of the standard diagnostic criteria of AUD are associated with genetic, neurophysiological, behavioural, and social factors. Some phenomenological descriptions and analyses highlight that alcohol-dependent patients typically experience powerlessness, suffering and self-stigma, and these experiences routinely accompany their drinking habits. For many patients, alcohol drinking is constitutive of their form of life. Undoing alcoholism would require "redoing oneself," by acquiring new goals and expectations, and developing novel habits that may give a new meaning to one's lived experiences. Let's now examine how different theoretical frameworks can fruitfully integrate these features of AUD.

3. Explanatory integration beyond reduction

[3] Like one of Smith's (1998) participants says: "I thought if I tried just drinking half of this bottle today, and that's the half for the next day. I painted a big thick line on the bottle. It never worked because the second you got down to that line you said, well I might as well just finish it" (218).

Integrating two or more accounts of a phenomenon consists in combining the concepts, evidence, results, or methods involved in those accounts into one integral explanatory account. The resulting integrated account explains different aspects of the phenomenon, displaying how such aspects are logically, probabilistically, constitutively, or causally related. To the extent an explanatory account is integrated, it yields understanding of the phenomenon as a multifaceted whole.

An integrated explanatory account of some target phenomenon requires some kind of dependence (e.g., logical, probabilistic, constitutive, or causal dependence) between the accounts to be integrated. If two accounts are fully independent, and they do not display any logical, statistical, evidential or conceptual dependence, then they cannot be integrated in an explanatory account. If they are fully independent, the two accounts are mutually irrelevant, and each would be unconstrained by the evidence, concepts, results or methods, on which the other account relies. Two accounts are independent to the extent they each enjoy many epistemic autonomies to a great degree. An account of a given phenomenon can enjoy different kinds of *epistemic autonomies* with respect to another account of the same phenomenon. Specifically,

(i) autonomy in the selection and use of taxonomic categories
(ii) autonomy in the selection of theoretical vocabulary
(iii) autonomy in the choice of methods of investigation
(iv) autonomy in the selection of and weight given to relevant evidence

If an account does not enjoy any of these autonomies with respect to another, then the relationship between the two accounts is one of full dependence.

Full dependence between accounts in the science of mind and brain is generally understood in terms of a "classical" notion of *reduction*. Different strategies for scientific reduction have been developed in the philosophy of science (e.g., Nagel 1961; Schaffner 1993; Bickle 2006). Although these strategies aim at establishing that a certain phenomenon (entity, property, or process) is identical to or fully explained by another, more basic phenomenon (entity, property, or process), or that a certain account (concept, model, or theory) can be logically derived from another more basic account, they all share at least the assumptions that phenomena and their scientific study belong in different "levels", and that the concepts of a reduced explanatory approach should be connectable to the concepts of the reducing explanatory approach.

While it's often claimed that "reductionism has dominated both research directions and funding policies in clinical psychology and psychiatry" (Borsboom, Cramer, & Kalis 2018),[4] actual attempts at reducing particular psychiatric phenomena to lower-level neural, molecular, or genetic underpinnings are sparse, "patchy" and "partial" (cf., Schaffner 2013). As Schaffner (2013, 1018) explains, "[i]n the past fifty years, a reductionistic approach in the biomedical sciences and in psychology has become far less imperialistic and considerably more fragmented and tentative". So,

[4] Borsboom, Cramer and Kalis (2018) argue that if a symptom network modelling approach to understanding and treating mental maladies is the correct approach, then reductionism in psychiatry is false. One of the assumptions of this argument is that a network approach is incompatible with a causal modelling approach aimed at inferring common (latent) causes of observed correlations between symptoms. Their argument seems then to equate *reduction* of a set of symptoms of a mental malady to a set of neurobiological structures and *inference* of a "latent" cause of a set of symptoms of a mental malady. But this is confusing, since the network approach, which Borsboom, Cramer and Kalis (2018) advertise, is compatible with a causal modelling approach aimed at inferring common (latent) causes of observed correlations between symptoms (Bringmann & Eronen 2018). Furthermore, reduction relationships typically hold between models or theories, and successfully inferring to a latent cause of a symptom does not entail the symptom has thereby been reduced to that cause.

because reductionism does not obviously offer an adequate framework for integrating different accounts in psychiatry in a way that comports with successful inter-field explanatory practices, we leave it on the side.

3.1 Dimensional computational phenotypes

Lying in between full independence and dependence, there are several intermediate positions, which involve relationships of partial dependence or mutual constraint (Kaplan 2017). One increasingly popular framework in psychiatry (Montague et al. 2012; Heinz 2017) appeals to David Marr's (1982) three-level framework for analysing information-processing systems. The *computational level* specifies what input-output function the system computes and why that type of system ought to compute that function. The *algorithmic level* specifies the effective procedures and representations employed by the system. The *implementation level* specifies how those procedures and representations are physically realized in the system.

Reinforcement Learning (RL) approaches to computational modelling have been blooming in psychiatry in the last few years (Maia & Frank 2011; Adams, Huys & Roiser 2016). Within RL, the computational level specifies the problem of learning what to do in an unfamiliar environment so as to maximize a numerical reward signal (Sutton & Barto 1998). Model-free control and model-based control are two families of RL algorithms that can be used to solve this problem. Model-based control algorithms learn a model of the environment, which they use to compute the expected value of possible actions by simulating their consequences. Model-based control produces more accurate and flexible decisions than model-free control, but is also computationally expensive, since it requires the agent to simulate future possibilities. Model-free algorithms do not exploit and search any model of the environment; they just store the long-run expected value of each action, computing them on-line, on the basis of a reward prediction error, the difference between predictions about the reward obtained by taking a certain action in a given state and the rewards actually received. Model-free control is less computationally costly than model-based control, but produces relatively inflexible decisions, which are similar to habits. If the agent's motivational state changes, or the structure of the environment changes, then the values "cached" by a model-free algorithm may be outdated and produce maladaptive choices. At the level of implementation, a wealth of neurobiological evidence suggests that the phasic activity of dopaminergic neurons in the basal ganglia encode prediction error signals that are recruited by the cortical-basal ganglia circuit for model-free (Montague, Dayan & Sejnowski 1996; Colombo 2014), and model-based control too (Langdon et al. 2018).

Relying on RL modelling, psychiatrists have started to identify possible *computational phenotypes* of mental maladies. Computational phenotypes are measurable behavioural, psychological and neural types defined in terms of specific parameters extracted from specific computational models of a given task on the basis of behavioural, psychological, and neurophysiological data (Montague, Dolan, Friston, & Dayan 2012; Patzelt, Hartley & Gershman 2018). As we'll explain below in Section 4.3, one clinically relevant[5] computational phenotype is a parameter that controls the trade-

[5] What's clinically relevant is a function of the "disease," "illness," and "sickness" aspects of a possible mental malady (Heinz 2017, 6). "Disease" refers to a biological, or psychological abnormality that is causally implicated in maladaptive behaviour, such as dampened dopaminergic firings, ineffective reward-based learning, and memory impairment in alcohol-dependent patients. "Illness" refers to the subjective experience of a malady, such as a sense of anxiety and bodily suffering in AUD. And "sickness" refers to impairment in social participation, where a person may be unable to learn and comply with local social norms, to communicate or interact smoothly with other people.

off between model-based and model-free processes in humans (Daw et al. 2011).[6] A range of psychiatric symptoms central to both AUD and many compulsive disorders has been associated with values of this phenotype, where model-based control is reduced in favour of model-free control (Gillian et al. 2016; Sebold et al. 2017; Voon et al. 2017).

Computational phenotypes can ground a dimensional framework for explanatory integration in psychiatry, which we'll display in Section 4.3 below. It is important to clarify already that, similarly to the dimensional approach taken by the Research Diagnostic Criteria (RDoC) initiative of the National Institute of Mental Health (National Institute of Mental Health 2010), dimensional approaches like ours assert that mental maladies should be understood as quantitatively, rather than qualitatively different from non-pathological psychological functions. Unlike the RDoC, our proposal does not assume that all mental maladies must have a localizable neurophysiological correlate; their organic correlate might be widely distributed and have diffuse effects at multiple spatial and temporal scales. And, unlike the RDoC, we do not subscribe to the idea that different levels of psychological function should be defined on the basis of genetic or neurophysiological dysfunction (Insel et al. 2010, 749). We propose instead that different levels of a psychological function should be defined more abstractly, in terms of different levels of a computational phenotype, and that mental maladies should be conceived as regions of the mathematical space defined by a set of clinically relevant computational phenotypes (e.g., balance between model-based and model-free control, delay discounting, learning rate, sensitivity to other agents' mental states).

In summary, our dimensional framework conceives of mental maladies as regions of the space defined by the computational phenotypes, understands levels as Marr's levels of analysis of a target computing system, and pursues explanatory integration by uncovering the common computational structure of apparently different maladies.

3.2 Mechanisms

Though Marr (1982, 25) claimed "the three levels are only rather loosely related" and emphasised the top-level as "critically important from the information-processing point of view" (27), in fact each one of Marr's three levels places taxonomic, theoretical, and evidential constraints on the other two levels of analysis (Colombo 2015, Sec. 4). Paying special attention to the implementation (or mechanistic) level, some have argued that Marr's levels "are just different aspects of the same mechanistic explanation" (Piccinini & Craver 2011, 303), and that explanatory integration in the mind and brain sciences should be grounded in the notion of a *mechanism* (Craver 2007).

Within the *mechanistic framework*, explanatory integration proceeds by revealing multi-level mechanisms responsible for phenomena. Mechanistic levels are levels of organization (not analysis), and are not individuated on the basis of considerations concerning scientific representation. The mechanism and its causal activities are at a higher level than the mechanism's constitutive component parts and operations; and, in turn, the mechanism's component parts and operations are at a higher level than their sub-components. The relationship between mechanistic levels is one of physical constitution, not causation; and talk of levels in this framework refers to part-whole relationships within mechanisms.

Mechanistic integration of different levels proceed by decomposing a system believed to be responsible for a phenomenon into its functionally relevant components, and by localizing which function is performed by which physical component when the mechanism produces the

[6] Specifically, this computational phenotype corresponds to the parameter ω in the following component of a hybrid, model-based and model-free algorithm for computing the Q-value of state and action pairs: $Q(s, a) = \omega Q_{MB}(s, a) + (1-\omega)Q_{MF}(s, a)$ (Daw et al. 2011, Supplemental Experimental Procedures).

phenomenon (Bechtel & Richardson 2010). According to this approach, integrating different mechanistic levels of explanation consists in decomposing, localizing, and recomposing a mechanism with the aim of displaying how entities and operations at many different levels are related to one another and contribute to the production of the target phenomenon to be explained.[7]

Within the mechanistic framework, Kendler, Zachar & Craver (2011) individuate mental maladies with *mechanistic property clusters*, that is: clusters of properties underlain, produced, supported or maintained by a mechanism. This view entails that explanatory integration of computational accounts and phenomenological analyses is successful to the extent such accounts can each reveal structures that produce, underlie and maintain a mental malady (Kendler 2008; Murphy 2013). Kaplan (2011, 347) captures this commitment in terms of a "model-mechanism-mapping constraint," whereby a model of a phenomenon has explanatory power to the extent that: "(a) the variables in the model correspond to identifiable components, activities, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these (perhaps mathematical) variables in the model correspond to causal relations among the components of the target mechanism." Compliance with this constraint would guarantee that different accounts of a target mental malady combine concepts, results, and methods to uncover a single mechanism responsible for the malady.

In summary, Kendler, Zachar & Craver's (2011) mechanistic framework conceives of mental maladies as property cluster mechanisms, understands levels as levels of physical organization within a mechanism, and pursues explanatory integration by combining concepts, results and methods from different fields in the service of discovering the mechanism responsible for a malady.

3.3 Networks of symptoms

Borsboom & Cramer's (2013) *network of symptoms* framework is another prominent approach for explanatory integration in psychiatry. Unlike the mechanistic framework, the network of symptoms framework understands psychiatric maladies as *alternative, stable states of networks of strongly connected symptoms*. These networks of symptoms need not have a common mechanism that's causing them. Symptoms here are not indicators of some underlying condition that causes them, but are understood as interconnected variables that are constitutive of mental maladies. The network of symptoms framework pursues explanatory integration in psychiatry by constructing networks of symptoms that reflect interdependencies between various neurobiological, psychological, behavioural, social and cultural symptoms (Borsboom, Cramer, & Kalis 2018).

Symptoms can be activated by external conditions, for example by the presence of empty bottles of beer in the environment; they can also be triggered by internal states, such as when steroids interfere with synaptic plasticity to impair long term potentiation in the hippocampus causes a "blackout" (Wetherill & Fromme 2016). As the network strategy understands symptoms as statistically and causally connected variables that can change over time, activation of some symptom can cause activation (or suppression) of some other symptom. When certain symptoms in a network are co-active and have the appropriate causal strength, a mental malady emerges.

In summary, Borsboom & Cramer's (2013) network of symptoms framework conceives of mental maladies as stable, interacting symptoms *without* an underlying common cause, posits no levels of organization where different symptoms would lie, and pursues explanatory integration by

[7] While the mechanistic framework does not entail a commitment to either reductionism or anti-reductionism, many mechanists concerned with explanatory integration have criticized the idea that reduction should be understood as a relationship between theories or models, emphasising the importance of multilevel explanations grounded in the pursuit of mechanism discovery (Darden 2006).

constructing networks of symptoms studied in different fields and by holistically capturing the structure of the relationships between these symptoms.

4. How to pursue explanatory integration in psychiatry

We now examine some of the theoretical and practical virtues of the three frameworks for explanatory integration in psychiatry we have outlined, as well as their limitations. Our overall conclusion is that our dimensional framework grounded in the notion of computational phenotype is the best for effectively integrating computational and mechanistic explanations with phenomenological analyses of mental maladies, in a way that comports with successful practices in psychiatry.

4.1 Mechanistic integration of computation and phenomenology

Kendler, Zachar & Craver's (2011) mechanistic framework has several attractions. It denies that mental maladies can be adequately understood as natural kinds defined in terms of necessary and sufficient conditions. They argue mental maladies should instead be understood as mechanistic property clusters. These clusters consist of sets of varying symptoms that are produced, stabilized, and maintained by some mechanism. Mechanisms individuate mental maladies, and different properties of a mental malady—say, biological, psychological and behavioural properties—would be properties of a single mechanism, at different levels of organization.

Although we are "far from being able to define plausible stability-producing mechanisms for most psychiatric disorders" (Kendler, Zachar & Craver 2011, 1148), one goal of current psychiatric practice is to discover and localize multiple, causal factors at different spatial and temporal scales that might constitute the mechanism of AUD. Such factors make a difference to whether a person compulsively seeks and takes alcohol, loses control in limiting alcohol intake, and tends to have negative emotional states associated with craving for alcohol and withdrawal. The mechanistic framework does not privilege any particular level of organization. The level of brain physiology, for example, is implicated in AUD, where alcoholic patients show decreased level of dopaminergic signalling in the ventral striatum, and increase of the corticotropin-releasing factor in the amygdala (e.g., Chen et al. 2011).

Despite these virtues, the mechanistic framework shows some theoretical and practical limitations in helping psychiatrists to pursue explanatory integration of mechanisms with computational and phenomenological accounts. Consider phenomenological methods. One of their common aims is to elucidate the structure of subjective intentional experiences, which cannot be detached from the whole circumstance of an embodied, ecologically situated patient. So, one of their aims is to provide patients, psychiatrists or therapists with a holistic account of the structure of the experience involved in a mental malady; and the pursuit of this aim plays distinctive roles in successful practices of diagnosis, sense-making and therapy (Parnas & Sass 2008). Instead, mechanistic strategies like those suggested by Kendler, Zachar & Craver (2011) aim at explaining mental maladies in terms of the decomposable and localizable parts and operations of a mechanism that produces and maintains a cluster of symptoms. If the mechanistic strategies and phenomenological methods make inconsistent assumptions about mental maladies—the former assume that mental maladies can be spatially decomposed and localize, while the latter assume they cannot—then they enjoy a high degree of autonomy. To the extent the methods of phenomenology play a helpful role for successful diagnosis and therapy of mental maladies, the mechanistic framework is inadequate for integrating phenomenology in an overall account of a malady of interest.

This leads us to the second limitation of Kendler, Zachar & Craver's (2011) mechanistic framework: at best, within a mechanistic framework, conscious experiences contribute to put into focus what need to be explained with the vocabulary and taxonomies of sciences aiming at

discovering the mechanisms of mental maladies. Phenomenological taxonomies—for example, subjective descriptions of the "ontological insecurity" experienced by alcohol-dependent patients—can be considered subjective, personal-level descriptions of pathological, sub-personal mechanisms giving rise to cravings (cf., Colombo 2013 on addiction and the personal/sub-personal distinction). These descriptions can contribute to elucidate the nature of the malady to be explained, but they do not provide a constraint on the adequacy of mechanistic accounts, which ultimately appeal to different taxonomies from genetics, neuroscience, psychology, and ethology. For, generally, phenomenological descriptions do not reliably map onto the types of entities and processes posited by mechanistic accounts. For example, some alcohol-dependent patients experience of time and temporal relationships as "circular"—as long as drinking continues, the future is just a re-enactment of the past, and future outcomes are just as valuable as past ones (Thune 1977)—but this experienced circularity does not map in any meaningful way onto the temporal relationships displayed by their neural processes. Because phenomenological descriptions and mechanistic accounts enjoy a relatively high degree of taxonomic autonomy, the mechanistic framework is theoretically inadequate for pursuing explanatory integration of mechanisms with phenomenological analyses.

If, within Kendler, Zachar & Craver's (2011) mechanistic framework, phenomenology enjoys too high a degree of methodological and conceptual autonomy, computational accounts enjoy *too little* autonomy, in a way that does not comport with successful practices in psychiatry and other sciences of mind and brain. Computational models are conceived of as "elliptical or incomplete mechanistic explanations" within the mechanistic framework (Piccinini & Craver 2011, 284). Mechanist philosophers focus their attention on the level of biological implementation, on which the explanatory value of accounts at Marr's algorithmic and computational levels would depend. But successful practices in psychiatry as well as in other sciences of mind and brain show that computational models need not map onto biological mechanisms to be practically useful in a variety of clinical and research contexts, to answer counterfactual questions and unify different phenomena under the same computational description (see, e.g., Chirimuuta 2018; Weiskopf 2018). This suggests that computational models should not be understood as mechanistic sketches: their explanatory value does not completely depend on their capacity to uncover mechanisms; it also suggests that the taxonomies employed at the computational and algorithmic levels of analysis should enjoy "soft" constraints with respect to the details of physical implementation, as computational analyses and algorithmic models do not make any claim about the spatial localization and organization of the components they posit, which may be implemented in multiple physical mechanisms (Elber-Dorozko & Shagrir 2019).

Voon et al. (2017) offers an example of the explanatory and taxonomic autonomies computational models should enjoy with respect to mechanism. Voon and collaborators review several lines of evidence that indicate the clinical and translational relevance of RL model-based control across multiple psychiatric disorders with different underlying mechanisms, including binge eating disorder, obsessive compulsive disorder, and AUD. Specifically, self-reported severity of alcohol use has been found to be associated with impairments in model-based control (Gillan et al. 2016), treatment outcome abstinence duration (Voon et al. 2015); the interaction between reduced model-based control and high expectations about the positive effects of alcohol has been found to predict risk of relapse (Sebold et al. 2017). Furthermore, computational modelling of the balance between model-free and model-based control provides a theoretical foundation for therapeutic interventions that aim to increase model-based control and inhibit model-free processes underlying temptation

and societal pressure, such as training at cognitive bias modification[8], which has been found to improve treatment outcome (Wiers et al. 2011; Heinz et al. 2017; see also Moutoussis et al. 2018).

This body of evidence indicates that a specific computational phenotype—*viz*. balance between model-based and model-free control in learning tasks—can unify apparently different compulsive disorders. It also indicates that computational models can support some counterfactual predictions about treatment outcome and risk of relapse. While model-based control has been associated with ventromedial prefrontal cortex and caudate activity, and model-based computation shares a dopaminergic foundation with model-free control (Deserno et al. 2015), the success of computational phenotypes in unifying apparently different disorders and supporting counterfactual predictions does not obviously depend on their mapping onto specific neural structures. If Kendler, Zachar & Craver's (2011) mechanistic framework does not acknowledge the relative degree of explanatory and taxonomic autonomies of computational models, then it cannot adequately integrate computational accounts in explanatory accounts psychiatry.

4.2 Network integration of computation and phenomenology

Borsboom & Cramer's (2013) network of symptoms framework presents many attractive features too. It eschews ill-defined talk of levels, as well as the simplistic identification of mental maladies with neurobiological states. Although the covariance between symptoms in a network can warrant causal conclusions about a target mental malady, and can thus uncover variables for intervention, these conclusions don't assume the covariance between symptoms arises from some common latent causes (Borsboom, Cramer & Kalis 2018). In this sense, Borsboom & Cramer's (2013) network of symptoms framework is "flat": it does not seek to uncover causes of symptoms at different levels of organization. Different features of a mental malady—say, craving, positive expectations about the outcomes of drinking, or reduced dopaminergic firing—are integrated as different interconnected nodes in a "flat" network structure. This kind of integration highlights statistical (or causal) patterns of heterogeneous variables that characterise a mental malady, while it can offer plausible accounts of comorbidity, in a way that comports with the more holistic methods of phenomenology.

Despite its attractions, Borsboom & Cramer's (2013) network of symptoms framework doesn't adequately integrate computational accounts and phenomenology. Within this framework, computational phenotypes, mechanisms, and phenomenological analyses enjoy a relatively high degree of conceptual, evidential and methodological autonomy. Consider phenomenological descriptions. Though Borsboom and Cramer (2013) and Borsboom, Cramer & Kalis (2018) do not address this point, phenomenology might at best constrain networks of symptoms indirectly. In fact, Borsboom, Cramer & Kalis (2018) argue that the covariation between symptoms in a network "can be seen to *make sense*" (20), because symptoms often correspond to intentional mental states, that is, to mental states that are about something. For example, the desire to drink alcohol is about drinking alcohol. Since intentional mental states display "a rational relation," Borsboom and colleagues suggest that networks of symptoms allow us to understand the lived experience involved in a malady. While this suggestion is a promising start, it falls short of providing a convincing answer to the question of how phenomenological descriptions or analyses fit or constrain a network of symptoms structure. After all, many phenomenological analyses of AUD experiences are focused on "pre-intentional" mental states like moods that need not be about any specific object in the world. Moods can be understood as providing subjects of experience with a background sense that

[8] This type of training is based on computer tasks performed with a joystick. The joystick is used to push alcohol-related images on the screen away and to pull images of water and alcohol-free beverages closer. When an image is pushed away, it becomes smaller; when it is pulled closer, it becomes larger. Alcohol-dependent patients taking this training in addition to normal behavioural therapy have a lower chance of relapse in comparison to patients who don't undergo this training.

structures their engagements with the environment infusing them with meaning (Heidegger 1962, 176; Jaspers 1913/1997, 688ff).

Even if a phenomenological description or analysis of AUD included only intentional states, it's far from obvious how their "rational" relationships within a network of symptoms should be specified. Specifying them in terms of logical or semantic relationships might be a promising route, but it's likely these relationships may not always track the causal relationships uncovered by network analysis. For example, some phenomenological description like Smith's (1998) and Flanagan's (2013) highlight that several alcohol-dependent patients truly believe that drinking will not make their suffering disappear and genuinely desire to stop drinking; yet, an overwhelming majority of alcohol-dependent patients will relapse within their first year of sobriety (Beck et al. 2012). Unless a network of symptoms include other mental states that could explain and rationalize the apparent inconsistency between relapse and the conscious belief that one should remain abstinent and desire that one remains abstinent, rational and causal relationships in the network will present a mismatch. And this mismatch will not promote "sense-making" of the lived experiences involved in AUD.

The network of symptoms framework can include nodes corresponding to computational phenotypes. But, because Borsboom & Cramer's (2013) network of symptoms framework is meant to be "flat," it does not take account of the organizational relationships between the neurobiological components and causal activities that physically realize computational parameters and algorithmic transformations. Nor does the network of symptoms framework specify how the transformations posited at Marr's algorithmic level relate to what a system is computing and why the system is computing that function instead of another. So, this framework—to the extent it pits itself against latent variable (or common cause) models, which it need not do (Bringmann & Eronen 2018)—cannot integrate computational accounts of mental maladies.

Consider variables that correspond to entities and causal activities that realize a computational phenotype like the balance between model-based and model-free control. Some of these variables (e.g., level of dopamine release in the ventral striatum) are likely to be common causes of psychiatric symptoms (e.g., inflexible learning and craving for alcohol in certain environments). But one of the assumptions of Borsboom, Cramer & Kalis (2018) is that a network approach is incompatible with a causal modelling approach aimed at inferring common (latent) causes of observed correlations between symptoms; they claim: "If a network model is correct... there exists no common cause" (17). Despite this claim, however, network and latent variable models should *not* be seen as providing competing accounts, and instead should be considered as complementary strategies for understanding and treating mental maladies. Networks of symptoms can gain "depth" and allow for the integration of information about Marr's different levels of analysis, by using representations of networks that encompass latent variable structures (see e.g. Epskamp, Rhemtulla & Borsboom 2017), and that show how causal transactions between organized sets of variables systematically relate to computational transformations of information.

4.3 Dimensional integration of computation and phenomenology

Conceptualizing mental maladies in terms of dimensional computational phenotypes allows us to unify apparently different diagnostic categories on the basis of their common dimensional computational structure. Within a space of computational phenotypes, we can also answer counterfactual questions concerning how social behaviour, neural activity, and subjective experience would change, had the value of a certain computational phenotype defining that space changed. These are two reasons why computational phenotypes have explanatory power. Let us now put into focus the notion of a computational phenotype, and consider how a dimensional framework grounded in this notion can help psychiatrists pursue the integration of mechanistic and computational explanations with phenomenological accounts of mental maladies.

As we already mentioned, computational phenotypes are types of parameters defined within a computational model of a task (Montague, Dolan, Friston, & Dayan 2012; Patzelt, Hartley & Gershman 2018). Computational phenotypes include such model parameters as *rate of learning*, which controls the extent to which new information overrides old information, *delay discounting*, which determines the extent to which the present value of a reward is discounted with delay of its receipt, *loss aversion*, which controls the preference to avoid losses to acquiring equivalent gains, and *depth of reasoning*, which controls to what extent one considers the thoughts of other people in strategic reasoning.

Computational phenotypes are continuous parameters and define types of continuous (or dimensional) psychological functions. A set of computational phenotypes can be used to define an abstract space of human phenotypes, that is, a space of types of individuals who share modes of behaviour and information processing for a wide range of decision and learning scenarios. For example, AUD might correspond to a region of the space defined by *delay discounting*, *learning rate* and *trade-off between model-based and model-free control*. The choice of computational phenotypes most relevant to define a certain dimensional space for a target malady depends on evidence available about an individual's psychological and neurobiological dysfunctions, on the individual's level of social participation, and on the individual's affective life. It also depends on the practical clinical needs, and on clinicians' phenomenological insights into the condition of a patient.

Now, how does a space defined over dimensional computational phenotypes exactly promote explanatory integration in a way that is more theoretically adequate and practically useful than Kendler, Zachar & Craver's (2011) mechanistic framework and Borsboom & Cramer's (2013) network of symptoms framework?

Let's start from mechanism, clarifying the main differences between our own proposal and Kendler, Zachar & Craver's (2011). First, we do not assume an account of a mental malady is adequate to the extent it uncovers its neurobiological mechanism; unlike the mechanistic framework, we assume an integrative explanatory account should be judged in terms of practical success, not in terms of its ability to latch onto mechanisms that exist independently of human theorizing. Second, computational modelling enjoys a relative higher degree of autonomy within our framework; though it constrains and it is constrained by available mechanistic evidence, computational analyses and algorithmic models are not sketches of neurobiological mechanisms. Third and finally, and that, on our view, types of mental maladies are in part individuated on the basis of human classificatory practices, in particular practices involving computational and phenomenological analyses.

Computation and neural mechanisms can obviously be integrated within our dimensional framework. In keeping with Marr's three levels of analysis, computational phenotypes are realized by neurobiological mechanisms that transform exteroceptive, proprioceptive, or interoceptive inputs into behavioural, emotional, or cognitive outputs. Given that a computational phenotype such as *balance between model-based and model-free control* is extracted from a computational model on the basis of behavioural and neural data, different values of this phenotype will be associated with different environmental stimuli, but also different levels of activity in certain neural circuits in the medial prefrontal cortex, also in alcohol-dependent patients (Daw et al. 2011; Deserno et al. 2015; Sebold et al. 2017). More importantly, the trade-off between model-based and model-free processes towards model-free control can be a computational phenotype of a range of disorders underlain by different mechanisms but all involving compulsion or drug abuse (Gillan et al. 2016; Voon et al. 2017). In this way, dimensional computational phenotyping can ground a unified explanation of

several kinds of addictions beyond AUD, displaying their common computational structure within a certain space of computational phenotypes.

Consider phenomenology. Berrios and Marková (2013) argue that a dimensional approach to psychiatry is misguided and cannot integrate phenomenological analyses or descriptions. Their argument is that a dimensional approach to some phenomenon entails the possibility of measuring that phenomenon by concretely interacting with it. Because, according to Berrios and Marková, mental symptoms have "abstract" or "ideal attributes (meanings)," which cannot be measured, "mental symptoms can only be evaluated (not measured)" (78).

Berrios and Marková's (2013) argument is inconclusive. A dimensional approach to some phenomenon does not entail that that phenomenon must be measurable via interaction or must be a concrete object. Sets are not concrete objects; yet, their cardinality can be measured. In fact, measurement and dimensionality often involve the representation of ideal systems, such as the consumption of alcohol in the average household in a certain neighbourhood in a country. Furthermore, measurement theory is a heterogeneous field, where different authors with different epistemic commitments, understand the nature of the *relata* of an act of quantitative measurement differently. Regardless of the nature of the *relata* of measurements, Berrios and Marková's (2013) argument is at odds with the fruitfulness of psychometric and dimensional approaches to understanding mental maladies (cf., Hägele et al 2015; Heinz et al 2016). What's correct in Berrios and Marková's (2013) suggestion is that psychiatric research often involves phenomenological description, analysis, and interpretation of subjective experiences of suffering that cannot obviously be measured *only* with questionnaires, scales, experimental tasks, or bodily measurements.

Our dimensional computational framework, however, is responsive to phenomenological descriptions and analyses of mental maladies in two ways. First, Marr (1982, 22) says that "the most abstract is the level of *what* the device does and *why*". What a system does and why it does that instead of something else contribute to delineating the phenomenon to be explained (Shagrir 2010). Within our framework, phenomenological analyses and descriptions are charged with helping us specify what a system is meant to accomplish within a certain ecological context, in a way that demonstrates the aptness of what the system does in that ecology. For example, Laing's (1960) analysis of *ontological insecurity* displays psychoses as essentially bound up with one's sense of "ontological insecurity", where one feels they are losing their sense of self, reality, and meaningful social relationships. While ontological insecurity can usher in anxiety, withdrawal and avoidance, this concept can helpfully illuminate the phenomenon to be explained and its ecological constraints. A computational-level hypothesis informed by Laing's phenomenological analysis of ontological insecurity is that alcohol-dependent patients may fail to integrate afferent interoceptive and exteroceptive representations with self-referential representations. Couched in mathematical terms, this hypothesis can then be specified algorithmically, and tested in the light of behavioural and neural data.

Second and more generally, phenomenological analyses and descriptions can provide patients and clinicians with narrative *glue* that may help patients make sense of the relationship between their suffering and their computational phenotypes. The abstract, non-biological taxonomies of computational models can be more readily re-interpreted than mechanistic accounts in terms of phenomenological categories. These categories may help one see how different computational phenotypes might be related and may reflect one's lived experience of choices and perceptions of reality. They may help patients and their beloved answer "existential" questions about the point of the suffering involved in their malady (cf., Roberts 2000).

One important objection to our proposal is that rather than offering an alternative explanatory framework, what we are proposing just changes the topic: unlike the mechanistic and network of symptoms frameworks, our dimensional framework only re-defines the *explanandum* (i.e., mental maladies); it does not explain why or how mental maladies come about.

To address this objection, it's helpful to draw an analogy. Different quantities suffice to physically characterise a system. For example, if you want to characterize a spring undergoing simple harmonic motion, its mass, period, and the acceleration of gravity suffice (plus a constant k determined by Hooke's law). These quantities have dimensions. The dimension of the period of a pendulum is time [T]; the dimension of mass is [M]; the dimension of the acceleration of gravity is length divided by the square of a time, [L/T^2]; and the dimension of the constant k is [M/T^2]. If we want to know why the spring has a certain period of oscillation, then we can derive the dimensional structure of the spring by working out an equation that gives us one quantity of interest as a function of all the quantities on which that quantity depends. From knowledge of the dimensional structure of the spring, we can conclude that the period P is proportional to the square root of mass divided by k.

This type of dimensional analysis is commonplace in physics, *and* it allows us to find the functional relationships between a set of quantities. These functional relationships can provide us with information about why apparently different systems behave similarly by considering the common dimensional structure they share (Lange 2009). It also allows us to gain modal information about the behaviour of a system by allowing us to answer counterfactual questions about how change in some quantity of a system would influence change in some other quantity of that system. As Pexton (2014) puts it, this type of "[d]imensional explanation is not simply about reading off dimensions naturalistically from a system and combining them to get functional forms of dependence between variables. Rather it implicitly involves picking a conceptualisation of a target system that in part creates a perspective from which the dimensional architecture is constructed" (2350).

Now, in the current state of research in computational psychiatry, we are far from being able to specify a plausible set of computational phenotypes for most mental maladies. And computational phenotypes do not have any obvious dimension we are familiar with from physics; their dimensions need to be clarified within computational psychiatry and value theory. Yet, by using a dimensional structure defined by a set of computational phenotypes relevant to a target mental malady, we can not only represent *what* the malady might consist of. As we already pointed out, we can also see that apparently different mental maladies present a common dimensional computational structure. In this way, computational phenotyping can help psychiatrists understand what clusters of symptoms are produced by the same type of processes, and to what extent these processes are realized by common types of neurobiological mechanisms.

Furthermore, using a dimensional structure defined by a set of computational phenotypes allows us to gain clinically relevant, modal information about a mental malady. In particular, it can give us information about clinical heterogeneity and about possible targets for treatment. For example, Heinz et al. (2017) suggest that if impaired model-based control is a key computational phenotype of AUD that predicts relapse, then interventions aimed to enhance model-based vs. model-free control on the basis of behavioural and cognitive training, or pharmacological manipulations will be promising therapeutic strategies for treating AUD (see also Moutoussis et al. 2018). In this way, computational phenotyping involves models connecting change-relating variables that allow psychiatrists to answer counterfactual questions generated by an *explanandum* mental malady.

Conclusion

One of the aims of psychiatry is explanatory integration. How can different concepts, sources of evidence, and methods used in different fields be integrated to adequately explain why a certain mental malady emerges and how it can be effectively treated? In this paper, we have started to articulate a dimensional theoretical structure based on the notion of *computational phenotypes* of mental maladies to pursue explanatory integration in psychiatry. Examining the case of AUD, we have shown how our dimensional framework can structure the search for tailored treatments targeting patients' expectations, social environment, computational modes of control, and neurophysiology. Our proposal is compatible with attractive aspects of alternative frameworks for explanatory integration in psychiatry, like RDoC (Insel et al. 2010), Kendler, Zachar & Craver's (2011) mechanistic framework, and Borsboom & Cramer's (2013) network of symptoms frameworks; but, unlike these frameworks, our dimensional proposal allows us to more adequately integrate mechanism, computation and phenomenology in pursuing general explanatory accounts of mental maladies.

Acknowledgements
We thank Roberto Fumagalli, Alexander Genauck, Miriam Sebold, and two anonymous referees for this journal for helpful conversations and comments on previous versions of this paper. Work on this paper was financially supported by the Alexander von Humboldt Foundation and by the Deutsche Forschungsgesellschaft (grant DFG FOR 1617/2).

References

Adams, R. A., Huys, Q. J., & Roiser, J. P. (2016). Computational psychiatry: towards a mathematically informed understanding of mental illness. *J Neurol Neurosurg Psychiatry*, *87*(1), 53-63.

Bechtel, W. & Richardson, R.C. (2010/1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*, Second Edition. Cambridge, MA: MIT Press/Bradford Books.

Beck, A., Wustenberg, T., Genauck, A., Wrase, J., Schlagenhauf, F., Smolka, M. N., Mann, K., and Heinz, A. (2012). Effect of brain structure, brain function, and brain connectivity on relapse in alcohol-dependent patients. *Archives of general psychiatry*, 69(8):842–52.

Berrios, G. E., & Marková, I. S. (2013). Is the concept of "dimension" applicable to psychiatric objects?. *World Psychiatry*, *12*(1), 76-78.

Bickle, J. (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese*, 151, 411–34.

Borsboom, D., Cramer, A., & Kalis, A. (2018). Brain disorders? not really why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, 1–54.

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121.

Bringmann L. F., & Eronen, M. I. (2018). Don't blame the model: reconsidering the network approach to psychopathology. *Psychological Review*. 125:606–15.

Carel, H. (2011). Phenomenology and its application in medicine. *Theoretical medicine and bioethics*, *32*(1), 33-46.

Chen, G., Cuzon Carlson, V. C., Wang, J., Beck, A., Heinz, A., Ron, D., ... & Buck, K. J. (2011). Striatal involvement in human alcoholism and alcohol consumption, and withdrawal in animal models. *Alcoholism: Clinical and Experimental Research*, *35*(10), 1739-1748.

Chirimuuta, M. (2018). Explanation in computational neuroscience: Causal and non-causal. *The British Journal for the Philosophy of Science*. 69, 849–880.

Colombo, M. (2017). Experimental philosophy of explanation rising: The case for a plurality of concepts of explanation. *Cognitive science*, *41*(2), 503-517.

Colombo, M. (2015). For a Few Neurons More. Tractability and Neurally-Informed Economic Modelling. *The British Journal for Philosophy of Science*, 66, 713-736.

Colombo, M. (2014). Deep and Beautiful. The Reward Prediction Error Hypothesis of Dopamine. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 45, 57-67.

Colombo, M. (2013). Constitutive relevance and the personal/subpersonal distinction. *Philosophical Psychology*, *26*(4), 547-570.

Connor, J. P., Haber, P. S., & Hall, W. D. (2016). Alcohol use disorders. *The Lancet*, *387*(10022), 988-998.

Craver, C. F. (2007). *Explaining the brain*. Oxford: Oxford University Press.

Darden, L. (2006). *Reasoning in Biological Discoveries: Mechanism, Interfield Relations, and Anomaly Resolution*, New York: Cambridge University Press.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69, 1204–1215.

Deserno, L., Beck, A., Huys, Q. J., Lorenz, R. C., Buchert, R., Buchholz, H. G., Plotkin, M., Kumakara, Y., Cumming, P., Heinze, H. J., Grace, A. A., Rapp, M. A., Schlagenhauf, F., & Heinz, A. (2015). Chronic alcohol intake abolishes the relationship between dopamine synthesis capacity and learning signals in the ventral striatum. *European journal of neuroscience*, 41(4):477–86.

DSM-5 (2013). Arlington, VA: American Psychiatric Association

Elber-Dorozko, Lotem & Shagrir, O. (2019). Computation and Levels in the Cognitive and Neural Sciences. In M. Sprevak & M. Colombo (Eds) *Routledge Handbook of the Computational Mind*, pp. 205-222. Routledge.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network pschometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904-927.

Everitt, B. J., & Robbins, T. W. (2016). Drug addiction: updating actions to habits to compulsions ten years on. *Annual review of psychology*, *67*, 23-50.

Flanagan, O. (2013). Identity and addiction: what alcoholic memoirs teach. In: Fulford K, Davies M, Gipps R, Graham G, Sadler J, Strangellini G, et al., editors. , editors. The Oxford Handbook of Philosophy and Psychiatry. Oxford: Oxford University Press. pp. 865–88.

Fuchs, T. (2010). Phenomenology and psychopathology. In *Handbook of phenomenology and cognitive science* (pp. 546-573). Springer, Dordrecht.

Gallagher, S. (2004). Neurocognitive models of schizophrenia: a neurophenomenological critique. *Psychopathology*, *37*(1), 8-19.

Garbusow M, Schad DJ, Sebold M, Friedel E, Bernhardt N, Koch SP, Steinacher B, Kathmann N, Geurts DE, Sommer C, Müller DK, Nebe S, Paul S, Wittchen HU, Zimmermann US, Walter H, Smolka MN, Sterzer P, Rapp MA, Huys QJ, Schlagenhauf F, Heinz A (2016) Pavlovian-to-instrumental transfer effects in the nucleus accumbens relate to relapse in alcohol dependence. *Addict Biol* 21:719–731.
Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, 5, e11305.

Hägele, C., Schlagenhauf, F., Rapp, M., Sterzer, P., Beck, A., Bermpohl, F., ... & Heinz, A. (2015). Dimensional psychiatry: reward dysfunction and depressive mood across psychiatric disorders. *Psychopharmacology*, *232*(2), 331-341.

Heidegger, M. (1962). *Being and Time* (trans. Macquarrie, J. and Robinson, E.). Oxford: Blackwell.

Heinz, A. (2017). *A New Understanding of Mental Disorders: Computational Models for Dimensional Psychiatry*. MIT Press.

Heinz, A., Deserno, L., Zimmermann, U. S., Smolka, M. N., Beck, A., & Schlagenhauf, F. (2017). Targeted intervention: Computational approaches to elucidate and predict relapse in alcoholism. *Neuroimage*, *151*, 33-44.

Heinz, A., Schlagenhauf, F., Beck, A., & Wackerhagen, C. (2016). Dimensional psychiatry: mental disorders as dysfunctions of basic learning mechanisms. *Journal of Neural Transmission*, *123*(8), 809-821.

Heinz, A., Deserno, L., Zimmermann, U. S., Smolka, M. N., Beck, A., & Schlagenhauf, F. (2017). Targeted intervention: Computational approaches to elucidate and predict relapse in alcoholism. *Neuroimage*, *151*, 33-44.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., *et al* . (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry* , *167* (7), 748‑51 .

Jaspers, K. (1913/1997). *General Psychopathology*, 7th edition, J. Hoenig & M. W. Hamilton (trans.), Baltimore: Johns Hopkins University Press.

Kaplan, D. M. (2017). Integrating Mind and Brain Science: A Field Guide. In David Kaplan (ed.) *Explanation and Integration in Mind and Brain Science*. Oxford University Press.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183(3), 339-73.

Kendler, K. S. (2012). Levels of explanation in psychiatric and substance use disorders: implications for the development of an etiologically based nosology. *Molecular Psychiatry*, 17, 11-21.

Kendler, K. S. (2008). Explanatory models for psychiatric illness. *American Journal of Psychiatry* 165, 695–702.

Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders?. *Psychological medicine*, *41*(6), 1143-1150.

Kurth-Nelson, Z., O'Doherty, J. P., Barch, D. M., Denève, S., Durstewitz, D., Frank, M. J., Gordon, J. A., Mathew, S. J., Niv, Y., Ressler, K., & Tost, H. (2016). Computational Approaches for Studying Mechanisms of Psychiatric Disorders. In A. D. Redish and J. A. Gordon (Eds.) *Computational Psychiatry: New Perspectives on Mental Illness*. Cambridge, MA: MIT Press, pp. 77-99.

Laing, R. D. (1960). *The Divided Self: An Existential Study in Sanity and Madness*. Harmondsworth: Penguin.

Langdon, A. J., Sharpe, M. J., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions for dopamine. *Current Opinion in Neurobiology*, *49*, 1-7.

Lange, M. (2009). Dimensional explanations. *Nous*, *43*(4), 742–775.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, *14*(2), 154-62

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Merleau-Ponty M. (1945/1962) The phenomenology of perception (trans: Smith C). New York: Routledge & Kegan Paul.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72-80.

Moutoussis, M., Shahar, N., Hauser, T. U., & Dolan, R. J. (2018). Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Computational Psychiatry*, *2*, 50-73.

Murphy, D. (2013). The medical model and the philosophy of science. In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Sadler, G. Stanghellini, & T. Thornton (Eds.), Philosophy and psychiatry (pp. 966–986). Oxford: Oxford University Press.

National Institutes of Mental Health (2010). *From Discovery to Cure: Accelerating the Development of New and Personalized Interventions for Mental Illnesses*. *Report of the National Advisory Mental Health Council's Workshop: 31*. Bethesda, MD : National Institutes of Mental Health.

Parnas J, Sass LA. (2008) Varieties of "Phenomenology": on description, understanding, and explanation in psychiatry. In: Kendler K, Parnas J, eds. Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology. Baltimore, MD: Johns Hopkins University Press; 2008: 239–277.

Patzelt EH, Hartley CA, Gershman SJ. (2018) Computational Phenotyping: Using Models to Understand Individual Differences in Personality, Development, and Mental Illness. *Personality Neuroscience*. 1: e18, XX: 1-10. doi:10.1017/pen.2018.14

Pexton, M. (2014). How dimensional analysis can explain. *Synthese*, *191*(10), 2333-2351.

Piccinini, G. & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183(3), 283–311.

Powers, A. R., Bien, C., & Corlett, P. R. (2018). Aligning Computational Psychiatry With the Hearing Voices Movement: Hearing Their Voices. *JAMA psychiatry*. 75(6):640-641.

Roberts, G. A. (2000). Narrative and severe mental illness: what place do stories have in an evidence-based world?. *Advances in Psychiatric Treatment*, *6*(6), 432-441.

Sass, L. (2014). Explanation and description in phenomenological psychopathology. *Journal of Psychopathology*, *20*(4), 366-376.

Schaffner, K. (2013). Reduction and reductionism in psychiatry. In K. W. M. Fulford, D. Martin, G. T. G. Richard, G. George, Z. S. John, S. Giovanni, T. Tim, & F. S. Kenneth (Eds.), The Oxford handbook of philosophy and psychiatry. Oxford: Oxford University Press.

Schaffner, K. (1993). Theory Structure, Reduction and Disciplinary Integration in Biology. *Biology and Philosophy* 8(3): 319-347.

Sebold, M., Nebe, S., Garbusow, M., Guggenmos, M., Schad, D. J., Beck, A., ... & Heinz, A. (2017). When habits are dangerous: alcohol expectancies and habitual decision making predict relapse in alcohol dependence. *Biological psychiatry*, *82*(11), 847-856.

Sebold, M., Deserno, L., Nebe, S., Schad, D. J., Garbusow, M., Hägele, C., ... & Rapp, M. A. (2014). Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology*, *70*(2), 122-131.

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, 77, 477–500.

Shinebourne, P., & Smith, J. A. (2009). Alcohol and the self: An interpretative phenomenological analysis of the experience of addiction and its impact on the sense of self and identity. *Addiction Research & Theory*, 17(2), 152-167.

Smith, B. A. (1998). The problem drinker's lived experience of suffering: an exploration using hermeneutic phenomenology. *Journal of Advanced Nursing*, *27*(1), 213-222.

Smith, D. W. (2018). Phenomenology. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/phenomenology/>.

Sprevak, M. & Colombo, M. (Eds.) (2019). *Routledge Handbook of the Computational Mind*, London: Routledge.

Stephan, K. E., Binder, E. B., Breakspear, M., Dayan, P., Johnstone, E. C., Meyer-Lindenberg, A., ... & Flint, J. (2016). Charting the landscape of priority problems in psychiatry, part 2: pathogenesis and aetiology. *The Lancet Psychiatry*, *3*(1), 84-90.

Tabakoff, B., & Hoffman, P. L. (2013). The neurobiology of alcohol consumption and alcoholism: an integrative history. *Pharmacology Biochemistry and Behavior*, *113*, 20-37.

Thune, C. E. (1977). Alcoholism and the archetypal past: a phenomenological perspective on Alcoholics Anonymous. *Journal of Studies on Alcohol*, *38*(1), 75-88.

Voon, V., Reiter, A., Sebold, M., & Groman, S. (2017). Model-based control in dimensional psychiatry. *Biological psychiatry*, *82*(6), 391-400.

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., ... & Robbins, T. W. (2015). Disorders of compulsivity: a common bias towards learning habits. *Molecular psychiatry*, *20*(3), 345-352.

Weiskopf, D. A. (2018). The explanatory autonomy of cognitive models. (2018). In David M. Kaplan (ed.), Explanation and Integration in Mind and Brain Science(pp 44–69). Oxford: Oxford University Press.

Wetherill, R. R., & Fromme, K. (2016). Alcohol-induced blackouts: A review of recent clinical research with practical implications and recommendations for future studies. *Alcoholism: clinical and experimental research*, *40*(5), 922-935.

Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychological science*, *22*(4), 490-497.

World Health Organization (2014). Global status report on alcohol and health 2014 (ISBN 978 92 4 069276 3). Geneva. Available at: www.who.int/substance_abuse/publications/global_alcohol_report/en/. Accessed July 7, 2018.