Erkenntnis
https://doi.org/10.1007/s10670-022-00533-x

ORIGINAL RESEARCH

Computational Modelling for Alcohol Use Disorder

Matteo Colombo[1]

Received: 18 January 2021 / Accepted: 10 March 2022

Abstract

In this paper, I examine Reinforcement Learning (RL) modelling practice in psychiatry, in the context of alcohol use disorders. I argue that the epistemic roles RL currently plays in the development of psychiatric classification and search for explanations of clinically relevant phenomena are best appreciated in terms of Chang's (2004) account of epistemic iteration, and by distinguishing mechanistic and aetiological modes of computational explanation.

Keywords Alcohol use disorders · Alcohol-avoidance training · Reinforcement learning · Computational modelling · Psychiatric classification · Psychiatric explanation

1 Introduction

In *An inquiry into the effects of ardent spirits upon the human body and mind*, American physician Benjamin Rush describes alcoholism[1] as an "odious disease" with

[1] 'Alcoholism' is the most commonly used term in English to refer to a kind of severe drinking problem that deserves clinical treatment. 'Alcohol dependence,' 'alcohol addiction,' and 'alcohol abuse' have also been used in clinical and non-clinical contexts with different connotations. Recent versions of diagnostic classifications like the Diagnostic and Statistical Manual of Mental Disorders (DSM-5, APA 2013) and International Classification of Diseases (ICD-11, WHO 2020) use the category 'alcohol use disorder.' Schomerus et al. (2011) review some of the evidence indicating that, *regardless* of how we refer to people with severe drinking problems, they are more likely to be held responsible for their condition, less likely to be regarded as mentally ill and tend to suffer more from social rejection and negative attitudes compared to people with other psychiatric diseases (see also Room 2005 on stigma and social inequality in alcohol use disorders).

✉ Matteo Colombo
m.colombo@uvt.nl

[1] Tilburg center for Logic, Ethics and Philosophy of Science Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Published online: 10 August 2022

noxious effects on body, mind and society (1811, 1). The “habitual use of ardent spirits”—writes Rush—causes “decay of appetite, sickness at stomach, and a puking of bile… [o]bstructions of the liver… [j]aundice and dropsy of the belly and limbs… [r]edness, and eruptions on different parts of the body… fetid breath… [e]pilepsy;” it also causes “[m]adness”, memory impairment, debilitation of the understanding and perversion of “the moral faculties” (Rush 1811, 5–7). Discussing several “remedies which are proper to prevent the recurrence of fits of drunkenness, and to destroy the desire for ardent spirits,” Rush recommends we learn to associate “the idea of ardent spirits, with a painful and disagreeable impression upon some part of the body,” for instance by learning to associate the sight of a bottle of rum with the memory of a disgusting event (1811, 28−9).

Fast forward to the twentieth century, and we find that contemporary diagnostic classifications of *alcohol use disorders* continue to be grounded in bodily, psychological and social symptoms associated with heavy drinking over time (Connor, Haber & Hall 2016). Some contemporary therapies are similar to Rush’s recommended treatments, too. For example, *approach bias modification* aims to change heavy drinkers’ approach tendencies towards consuming alcohol, by re-training associations between alcohol-related cues and seeking alcoholic beverages or paying a disproportionate amount of attention to them (Kakoschke, Kemps & Tiggemann 2017).

These similarities might suggest otherwise, but psychiatric classifications of alcoholism has changed significantly since Rush’s (1811) inquiry (Nathan, Conrad & Skinstad 2016). These changes have been accompanied by better understanding of the complex array of biological, psychological, economic, social and cultural risk factors for alcohol-use problems, and sharper awareness of their seriousness for individuals and public health (Carvalho et al. 2019; Rehm et al. 2010; Witkiewitz, Litten & Leggio 2019).

More recently, in an attempt to overcome some of the challenges faced by existing diagnostic schemes, including their limited specificity and stability (see e.g. Rehm et al. 2013),[2] to improve knowledge of the causal structure of psychiatric illnesses and to find effective treatments, psychiatrists have been increasingly relying on computational modelling (Huys et al. 2021; Maia & Frank 2011; Montague et al. 2012; Moutoussis et al. 2018; Seriés 2020; Colombo 2021).

Increased reliance on computational modelling in psychiatry raises several issues of philosophical, scientific and practical interest. Here, I examine Reinforcement Learning (RL) modelling practice in psychiatry, clarifying some of its roles in the development of psychiatric nosology and the search for explanations of clinically relevant phenomena. After describing some salient aspects of this practice (Sect. 2),

[2] Write Rehm & Room (2015, 1): “Modern diagnosis… claims objectivity and universality; diseases are supposed to be defined independently of the country or culture where the diagnosis is made. This premise might not hold true for alcohol use disorders. The current definitions rely on several criteria that are mostly consequences of heavy drinking over time, such as a persistent desire to cut down on alcohol consumption, withdrawal, and failure to meet expected social roles. Such symptoms have different social meanings in different cultures, which lead to surprising comparisons such as the observation that the prevalence of alcohol dependence in Latvia is more than 20 times higher than that in Italy, [though] Latvia and Italy have similar levels of alcohol consumption per person, and similar prevalence of liver cirrhosis and heavy drinking.”

I argue that the epistemic roles of RL approaches to psychiatry are best appreciated in terms of Hasok Chang's (2004; 2017) account of *epistemic iteration* (Sect. 3) and by distinguishing constitutive and aetiological modes of computational explanation (Sect. 4).

2 Drinking habits, alcohol-avoidance therapy and RL modelling

Alcohol Use Disorders (AUDs) involve the loss of control over alcohol intake, a malfunctioning "off switch" as Flanagan (2013) puts it. This loss of control typically causes failure to fulfil obligations at work or home, interpersonal or legal problems, negative thoughts, bad feeling and bodily pain when not drinking (Carvalho et al. 2019). Individuals suffering from AUDs are likely to persist in a pattern of self-destructive alcohol consumption, despite their explicitly stated desire to abstain and their knowledge of the noxious consequences of continuous drinking. In fact, the great majority of patients diagnosed with AUD have a relapse to uncontrolled alcohol use, craving and tolerance in the year after alcohol-use treatment (Batra et al. 2016; Brandon, Vidrine & Litvin 2007; Heinz et al. 2017; Moos & Moos 2006; World Health Organization 2018).

These observations are consistent with the hypothesis that AUDs—and addictions more generally—develop through a shift from flexible goal-directed behaviour to more rigid habitual behaviour in response to alcohol-related cues (Bechara 2005; Everitt & Robbins 2005).[3] In turn, the hypothesis that AUDs develop by an over-reliance on habits at the cost of goal-directed behaviour has motivated the study and therapeutic use of treatments like approach bias modification (Kakoschke, Kemps & Tiggemann 2017).

Alcohol-avoidance training is one version of approach bias modification (Wiers et al. 2011). In alcohol-avoidance training, participants are repeatedly shown images of alcohol-related cues or non-alcohol-related cues on a screen (e.g., pictures of alcoholic beverages vs. soft drinks). For each trial during a 15-minute session, participants use a joystick in order to pull towards themselves the non-alcohol-related cue or push away from themselves the alcohol-related cue just presented on the screen. Pulling the joystick increases the size of the image, while pushing the joystick decreases the size of the image, thus eliciting visual approach and avoidance effects, respectively.

Alcohol-avoidance training is quick, inexpensive and represents a promising therapeutic add-on for enhancing personalized treatment outcomes, and reducing relapse rates (for relevant evidence see Laurens et al. 2020; Wiers, Boffo, & Field 2018; Wiers, Van Dessel, & Köpetz 2020). However, despite its promise, the estimated effect sizes of behavioural therapies like alcohol-avoidance training are uncertain

[3] This idea resonates with personal memoires like Flanagan's (2013), according to which "for a certain kind of alcoholic, if, as is standard in philosophy of action, an action is defined as (mental intention/motive + behavior), then there comes a time in which none (or very few) of his actions are not constituted by some mental relation, frequently conscious or semiconscious, he has to alcohol… The *lebenswelt* [of an alcohol-dependent patient becomes] partly constituted by some relation (conscious seeking at one end, conscious suppressing at the other end) his action has to ethanol" (Flanagan 2013, 871).

and small (Cohen's $d=0.2$ to 0.3) (Boffo et al. 2019),[4] and it remains contentious the extent to which alcohol-avoidance training is efficacious against relapse, and AUDs more generally.[5] The contentious issue is whether reduction in goal-directed behaviour and increased reliance on contextually cued habits are relevant factors in the causal history leading up to individuals' AUDs or relapse into alcohol abuse.

RL modelling has proved itself to be helpful to address this issue. By fitting RL models to experimental data, one can probe possible algorithms generating observed behaviour, find neural correlates of computational variables and better understand the effects of therapeutic interventions. In particular, several studies have formalized the notions of habitual and goal-directed behaviour within the framework of RL, investigated individual differences in the balance between habits and goals, and made suggestions about how to optimize existing therapies, including alcohol-avoidance training (Voon et al. 2017; Gillan et al. 2016, Huys et al. 2021, 9–13).

RL is a computational approach to the problem of learning what to do based on the rewards and punishments produced by repeated interactions with the environment (Sutton & Barto 2018). Within RL, goal-directed action is typically formalized as model-based RL, while habits are formalized as model-free RL (Dayan & Balleine 2002; Daw et al. 2011; Dolan & Dayan 2013).

Model-based RL is a computational procedure (i.e., an algorithm) for solving reinforcement learning tasks that acquires and uses a representation of the dynamics of the environment. Specifically, it recruits a decision tree specifying the consequences of different actions in different states of the environment and their goodness. Agents implementing model-based RL pursue their goals by searching this decision tree, simulating the outcomes of potential actions and choosing the best action (i.e., the action producing the outcome with the highest expected reward).

For example, if your goal is abstinence, then, by implementing model-based RL, you will simulate the outcomes of different actions you may take in various situations, and make choices based on their expected consequences in relation to your goal. Importantly, model-based RL algorithms are sensitive to changes in both motivation and the relationship between different events in the environment. For example, by relying on model-based RL, if you are not thirsty, have nausea or know the pub is closed now, you will probably not go for a beer or will not enjoy it.

[4] For comparison, popular contemporary pharmacological treatments for AUDs such as disulfiram, acamprosate and naltrexone have small to medium estimated effect sizes (Cohen's d around 0.5) (Witkiewitz, Litten, & Leggio 2019).

[5] Again for comparison, Montague et al. (2012, 72) motivate computational approaches to psychiatry by highlighting "the (almost) unreasonable effectiveness of *psychotropic medication*. These medications are of great benefit to a substantial number of patients; however, our understanding of why they work on mental function remains rudimentary." According to Montague and collaborators, computational modelling helps researchers to enhance their understanding of why drugs work, because the language of computation offers "appropriate intermediate levels of description that bind ideas articulated at the molecular level to those expressed at the level of descriptive clinical entities" such as AUDs." (cf. Colombo & Heinz 2019). In the philosophy of psychiatry, Tsou (2012) offers an account of pharmacological drugs as experimental tools for uncovering neurobiological mechanisms of psychiatric disorders like schizophrenia. My focus on a behavioural therapy for AUDs and the role of computational modelling in uncovering casually relevant factors of relapse is meant to complement accounts like Tsou (2012) focused on pharmacology and mechanism.

Model-free RL is a computational procedure that does not acquire or use a model of the environment. Agents implementing model-free RL make choices using cached information about the goodness of states (or state-action pairs) acquired and updated gradually, based on trial-and-error interactions with the environment. Unlike model-based RL, the cached information in model-free RL—say, the expected reward associated with drinking another beer—is not immediately sensitive to changes in motivation or in the environment.

For example, based on stored associations between, say, the clinking sound of a bottle and the goodness of consuming a beer, an alcohol-related cue in the environment, like that clinking sound or an advertisement, will likely trigger approach behaviour towards drinking *despite*, and not because, of the actual consequences of this behaviour. This relative insensitivity to outcome de-evaluation and environmental change is the mark of habitual behaviour. And several computational investigations have examined whether it is also the mark of AUDs, and addiction more generally.

Sebold et al. (2017), for example, wanted to better understand the role of model-based vs. model-free RL in AUD and associated treatment outcomes. Specifically, they wanted to clarify whether and how patients' goals, habits, environmental cues and subjective expectations about the reward value of drinking an alcoholic beverage are causally relevant for understanding AUDs and particularly for patients' treatment outcomes.

Sebold et al. (2017) had experimental participants, including patients diagnosed with AUD and healthy controls, to complete a two-step task while undergoing brain imaging. Individual participants were subsequently contacted for a personal assessment of their drinking behaviour at regular intervals of time over a period of one year.

As the two-step task is a sequential decision-making task that allows researchers to disentangle the causal contributions of habits and goals formalized as model-based and model-free RL respectively (Daw et al. 2011), Sebold et al. (2017) tested various computational models of their participants' behaviour in this task, ranging from "pure" model-free to "pure" model-based RL. Fitting these algorithmic models to trial-by-trial behavioural and neural data, they identified a "hybrid" algorithm concurrently including both model-free and model-based RL as the best fitting for all participants. One of the parameters of this hybrid model controls the trade-off between model-based and model-free RL for computing the expected value of state-action pairs and making choices. The lower the value of this parameter, the less influence does model-based control have on decision making compared to model-free control; and with reduced model-based control, choices in the two two-step task become relatively more inflexible, and are more likely to deviate from optimal behaviour (i.e., behaviour maximising expected cumulative reward in the task).

Sebold et al. (2017) found that reduced model-based control interacted with participants' expectations for the high reward value of drinking alcohol to explain observed variance in individual participants' risk of relapse. While this finding puts pressure on a simplistic understanding of addictions as merely "maladaptive habits," it supports the idea that any causal effect of reduced model-based control on one's risk of relapse probably depends on subjective expectations about the goodness of alcohol consumption. This, in turn, suggests that alcohol-avoidance training most likely works when paired with interventions targeted at patients' beliefs about the short- and long-term effects of alcohol on their own well-being.

With this example of RL modelling for AUDs in the background, let me now zoom in on two questions: first, what kind of impact should we plausibly expect from RL modelling on existing psychiatric classification? Second, does RL modelling in psychiatry enjoy explanatory power only to the extent it reveals the internal mechanistic structure of a phenomenon? I address these two questions in the next two sections respectively.

3 RL modelling, Classification and Epistemic Iteration

One of the results of Sebold et al.'s (2017) study is the identification of a *computational phenotype*. "A computational phenotype is a measurable behavioural or neural type defined in terms of some computational model" (Montague et al. 2012, 72; see also Patzelt, Hartley & Gershman 2018). More precisely, the computational phenotype identified by Sebold et al. consists in a parameter within a hybrid model-based/model-free RL algorithm. This parameter might be called *balance between model-based and model-free control*, since it controls the trade-off between model-based and model-free RL in learning and decision making, and thus determines the computational state of an agent at a certain time and the transitions between different computational states of an agent at different times.[6] When the value of this parameter is low, the agent tends to make relatively inflexible ("habitual") choices in reinforcement learning tasks.

Sebold et al. (2017) provide some evidence that this parameter has a decent degree of variation both between and within individuals, and within the same group of individuals, too. So, it may be used to distinguish different types of agents, say, those with a *balanced* vs. *imbalanced* model-based/model-free control. More specifically, variation in *balance between model-based and model-free control* may help researchers to reliably distinguish groups of healthy individuals from those with psychiatric disorders—including schizophrenia (Culbreth et al. 2016) and obsessive-compulsive disorder, methamphetamine dependence and binge eating disorder (Voon et al., 2015)—and to track variation observed in relevant clusters of symptoms comprising compulsive behaviour and intrusive thought (Gillan et al. 2016). Variation in the computational phenotype *balance between model-based and model-free control* may also reliably predict, as well as contribute to explain, individual differences in treatment outcomes (Voon et al. 2017).

These observations lend some support to the idea that the computational phenotype *balance between model-based and model-free control* is potentially[7] relevant for various psychiatric purposes including the improvement of diagnostic classification;

[6] *Learning rate* (i.e., the extent to which ones' new information overrides old information), *delay discounting* (i.e., the extent to which one discounts the present reward value of an outcome with delay of its receipt) and *balance between exploration vs. exploitation* are other examples of computational phenotypes defined in terms of RL models.

[7] Although preliminary psychometric results indicate that *balance between model-based and model-free control* has good to excellent reliability, is stable within individuals and shows a good degree of inter-

but it remains unclear what kind of improvements in nosology RL is likely to stimulate and in what sense they would be "improvements."

To address these issues, it is important to highlight two properties of computational phenotypes, namely: they are *dimensional* and *trans-diagnostic* variables. That computational phenotypes are dimensional means that they provide us with representations of individuals' characteristics as points in a continuous parameter space, where there is no unique threshold between, say, "normal" alcoholic consumption and pathological drinking (Colombo & Heinz 2019). As we shall see in a moment, the success of computational phenotyping might thus refine, integrate and constrain dimensional approaches to nosology, and stimulate research within the Research Domain Criteria framework developed by US National Institute of Mental Health (RDoC) (Morris & Cuthbert 2012; Cuthbert 2014), which stand in contrast to classificatory systems like DSM that see disorders as either present or absent based on polythetic-categorical criteria.

That computational phenotypes are trans-diagnostic means that they are implicated across disorders or clinical outcomes. Because computational phenotypes are defined in terms of what they do within algorithmic models of perceptual, learning and decision making tasks, which are used for identifying causal factors, pathways and mechanisms common across existing diagnostic categories, computational phenotyping is likely to promote trans-diagnostic approaches (Dalgleish et al. 2020) and simplify existing diagnostic tests.[8]

If the practice of computational phenotyping coheres with and is likely to promote existing dimensional, trans-diagnostic approaches to classification, how would this constitute improvement in nosology? More generally, what's the right epistemology for analysing plausible roles of computational modelling in contemporary diagnostic classification? My suggestion is that we should think of the impact of computational phenotyping in psychiatry in terms of Hasok Chang's (2004; 2017) notion of *epistemic iteration*, where epistemic iteration is an historically situated process, "in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic [and practical] goals… What we have is a process in which we throw very imperfect ingredients together and manufacture something just a bit less imperfect" (Chang 2004, 226).

Chang (2004) reconstructs the history of thermometry—the measurement of temperature—from the 1600s in order to flesh out the notion of epistemic iteration and explain how measurement and classificatory standards could be established and improved without presuming or requiring one stable foundation indicating what the right standards should be, and against which scientific improvement should be assessed.

According to Chang (2004), the development of valid and reliable measurements of temperature started from humans' perceptual experiences of warmth and cold.

individual variability (Brown et al. 2020), it should be emphasised that we currently know little about its validity and test-retest reliability, which means its clinical relevance is uncertain.

[8] Witkiewitz, Litten, & Leggio (2019, 7) note: "a DSM-5 diagnosis of AUD requires 2 or more symptoms, out of 11, over the past year. That requirement equates to exactly 2048 potential symptom combinations that would meet the criteria of AUD." Because of this heterogeneity, where two individuals can have AUD without sharing a single symptom, it is also doubtful AUD in itself is a valid category.

Over time, systematic correlations between our perceptions of warmth/cold and the expansion/contraction of various materials in the environment were observed, and the stability of these correlations was probed with experiments and new technologies. Drawing on relatively stable correlations and new theoretical insight into the possible nature of temperature, various types of thermoscopes were designed in the 17th century, and "fixed points" such as such as the boiling and freezing points of water were identified. Accumulating theoretical, practical and technical knowledge ushered in the construction of quantitative scales of temperature, which, compared to thermoscopes, allowed for more informative comparisons between the outputs of different instruments. Such comparisons, along with further advancements in theory and instrumentation, stimulated refinements, helping scientists to distinguish between different concepts of temperature, discard some as inadequate, retain others, and calibrate novel, more reliable and valid thermometers.

From the point of view of epistemic iteration, scientific improvement (or progress) consists in achieving broadly agreed practical and/or epistemic aims. The actual aims of research communities along with their values set communal standards of appraisal for a given scientific project at a specific place and point in time. While improvement comes in degrees, standards of appraisal for a given system of classification and measurement change over time, due to changes in relevant aims and values, or under the pressures of disagreement, or the integration and competition of a plurality of approaches. An ongoing process of epistemic iteration can thus facilitate two modes of scientific progress, namely: "*enrichment*, in which the initially affirmed system is not negated but refined resulting in the enhancement of some of its epistemic virtues; and *self-correction*, in which the initially affirmed system is actually altered in its content as a result of inquiry based on itself" (Chang 2004, 228).

Revisiting the development of diagnostic classifications of alcoholism in the last 70 years (Nathan et al. 2016; Robinson & Adinoff 2016; Sellman et al. 2014), we find ongoing self-correction and enrichment in the search for relatively stable classifications that could facilitate particular epistemic and practical purposes, under the pressures of patients' needs, scientific disagreements about theory and therapy, bureaucratic and legal challenges, and a plurality of approaches to mental health problems.

The *Diagnostic and Statistical Manual of Mental Disorders* DSM-I (APA 1952) was the first standardized classificatory system developed in the United States of America. Compiled by a relatively small set of representatives of the US War Department, Veterans Administration and American Psychiatric Association working within a broadly psychodynamic and psychoanalytic theoretical framework (Grob 1991), DSM-I contains brief descriptions of diagnostic conditions, emphasising *reactions* to stressors in the environment as key causal factors of illness.

DSM-I classifies both alcoholism and drug addiction as *Sociopathic Personality Disturbances*, where "[i]ndividuals to be placed in this category are ill primarily in terms of society and with the prevailing cultural milieu, and not only in terms of personal discomfort and relations with other individuals" (APA 1952, 38). Biological factors were considered to be less important than personality and cultural factors for classifying and explaining alcoholism.

Various changes in theoretical and practical approaches to psychiatry took place in the 1960s and early 1970s, which ushered in DSM-II (APA 1968). While Jellinek's *The Disease Concept of Alcoholism* was published in 1960, new pharmacological and cognitive-behavioural treatments were being tested and evaluated. Because of these developments, the relatively small committee working on DSM-II paid more attention to biology; but, similarly to DSM-I, DSM-II continues to include brief descriptions of diagnostic conditions and conceive of alcoholism as a subcategory of *Personality Disorders and Certain Other Non-Psychotic Disorders*.

DSM-II explicitly defines *Alcoholism* as a category "for patients whose alcohol intake is great enough to damage their physical health, or their personal or social functioning, or when it has become a prerequisite to normal functioning" (APA 1968, 45). And three species of clinically relevant drinking problems are now distinguished, namely: *Episodic Excessive Drinking* (for individuals intoxicated four times per year), *Habitual Episodic Drinking* (for individuals intoxicated more than 12 times a year, or recognizably under the influence of alcohol more than once a week even though not intoxicated), and *Alcohol Addiction*, characterised in terms of an inability to abstain for one day, or constant heavy drinking for three months or more (APA 1968, 45). So, as portrayed by the DSM-II, alcoholism is a behavioural disorder along a continuum of excessive, habitual and compulsive drinking.

Psychometric work soon demonstrated the low levels of validity and reliability of DSM-II categories (e.g., Spitzer et al. 1978); and this work identified more reliable clusters of behavioural and biological signs and symptoms for classifying alcohol-dependent patients and explaining their compulsive drink-seeking behaviour, as well as phenomena like alcohol tolerance, withdrawal and relapse (e.g., Edwards and Gross 1976). These and other results in statistics and clinical studies plausibly influenced the compilation of DSM-III (APA 1980)—though the committee working on it took an explicitly a-theoretical, descriptive approach grounded in empirically validated, symptoms-based criteria, unlike the short descriptions offered by DSM-I and DSM-II.

DSM-III distinguishes *Substance Abuse*, which refers to substance uses with negative social or occupational consequences (including legal problems, which may arise, e.g., from car accidents due to intoxication), and *Substance Dependence*, which refers to substance uses involving tolerance and withdrawal. Importantly, DSM-III now separates the categories *Alcohol Organic Mental Disorders* and *Personality Disorders*, advising practitioners to use the expression "an individual with Alcohol Dependence" instead of "alcoholic" (APA 1980, 6). Thus, while alcoholism was neatly separated from disorders of personality, it was also divided into two distinct categories of alcohol use disorders, viz. abuse and dependence.

DSM-IV (APA 1994) introduced the category *Alcohol-related Disorders* and maintained the distinction between substance *abuse* and *dependence*, where "[d] ependence (on a substance) is defined as a cluster of three or more… symptoms occurring at any time in the same 12-month period" (Ibid., 176), instead of requiring either tolerance or withdrawal as in the DSM-III. However, as dissociations between compulsive substance use and tolerance/withdrawal were found (e.g., Saha, Chou & Grant 2006), those two categories of alcohol use disorders were unified into a single category of *Alcohol Use Disorder* in the fifth and most recent iteration of the DSM (APA 2013).

DSM-5 is still grounded in a broad consensus among psychiatrists about clusters of symptoms rather than evidence at multiple scales about the causes of psychiatric maladies; but in introducing the section *Substance-Related and Addictive Disorders*, DSM-5 makes now explicit reference to biological causes, mechanisms and pathways, and also to the reward system in the human cognitive architecture and its effects on learning, decision making and memory.

DSM-5 states that "[a]ll drugs [including alcohol] that are taken in excess have in common direct activation of the *brain reward system*, which is involved in the *reinforcement* of behaviors and the *production of memories*… The *pharmacological mechanisms* by which each class of drugs produces reward are different, but the drugs typically activate the system and produce feelings of pleasure, often referred to as a 'high.' Furthermore, individuals with *lower levels of self-control*, which may reflect impairments of brain inhibitory mechanisms, may be particularly disposed to develop substance use disorders" (APA. 2013, 481, emphases added). Not only does DSM-5 draw attention to the importance of brain-mediated factors, such as craving and reduced self-control for treatment and classification; but it also emphasises that substance use disorders and a "behavioural addiction" like gambling disorder might share common causal factors and might therefore be treated with similar interventions.

This potted history I just sketched of the classification of alcoholism across different editions of the DSM fits within Chang's (2004) framework. This history displays an iterative process of enrichment and self-correction, which starts and is guided by phenomenological observation, clinical experience, patients' needs and communal social, practical and epistemic values. Psychometric, psychological and neurophysiological studies uncover and probe correlations between phenomenological observations, risk factors, environmental cues, and behavioural and biological symptoms. "Fixed points" are implicitly or explicitly employed—such as, for example, a definition of substance use disorders grounded in heavy use over time (Rehm et al. 2013)—for triangulation and probing the reliability and validity of these correlations. Concurrently, pharmacological interventions contribute to uncover aspects of neural and genetic mechanisms and aetiological factors and pathways, and theoretical approaches compete, get abandoned or converge.

Through this co-evolutionary, pluralistic process, understanding of the causal structure of a disease like alcoholism acquires depth (for how this process might play out in research on addiction see Colombo 2013). With the rise of computational approaches like RL, we are now witnessing a rapprochement between diagnostic-classificatory projects like the DSM and frameworks for interdisciplinary research like RDoC, advertising the utility of trans-diagnostic, dimensional computational variables involved in reward-based learning and decision making for furthering certain classificatory, therapeutic and explanatory purposes.

Whether computational psychiatry will mark an improvement in nosology must obviously be judged by its fruits, not by its promises. But one thing we should recognize with Chang (2011) is that epistemic iteration is best pursued in tandem with a pragmatically motivated explanatory pluralism, according to which scientific progress requires a plurality of modes of explanation that are differentially assessed according to different norms of explanation without the requirement that they all tie into some fundamental norm of appraisal or ultimately converge on a single unified explanation. Underlying each mode are different vocabularies creating, shrinking or expanding ways of conceptualizing phenomena; and new conceptualizations invite theoretical competition, integration and co-evolution between classificatory and explanatory systems (Kellert et al., 2006: x; see Kendler 2012 for a similar view in psychiatry).

Having laid out a view of progress in psychiatric classification in terms of Chang's (2004) notion of epistemic iteration and sketched how computational phenotyping might contribute to this process, let me now make the connection to explanation and suggest that there is a plurality of modes of computational explanation in psychiatry, which vindicates the idea that epistemic iteration in psychiatry is best pursued with a commitment to explanatory pluralism.

4 RL modelling, and Constitutive and Aetiological Explanation

By 'explanation,' I refer to a "vehicle of representation" (e.g., speech, written text, diagrams, mathematical equations, etc.), which, given some epistemic or practical aim, bears information about some explanatory fact related to a phenomenon of interest. By 'computational explanation,' I mean a type of *explanans* vehicle (e.g., a RL model), which, given some epistemic or practical aim, researchers use to ascribe computational features and processes (i.e., rule-governed transformations of computational states) to a target system displaying some *explanandum* phenomenon.

Philosophers of science disagree about what explanatory facts are; but there is broad agreement that an *explanans* vehicle—for example, a computational model of a given decision-making task—has explanatory power over some *explanandum* at least to the extent some of the information it bears enables us to draw inferences about the counterfactual behaviour of the *explanandum*—for example, alcohol-dependent patients' treatment outcomes—so as to further certain communal epistemic or practical aims. In Woodward's (2003, 221) words: "the common element in many forms of explanation, both causal and noncausal, is that they must answer what-if-things-had-been-different questions."

There is also agreement that a prominent mode of causal explanation in the life sciences is mechanistic, where a mechanism is a spatially and temporally organized system of components and activities that collectively constitute some entity's doing something like, for example, neurons' generating action potentials, or having the capacity to do something like, for example, humans' capacity to make goal-directed decisions (Craver & Tabery 2015).

The disagreement concerns the scope of mechanistic explanation and its relation to computational modelling, where some philosophers argue that any computational model has explanatory power only to the extent it uncovers the mechanism responsible for the *explanandum* (e.g., Kaplan 2011; Piccinini & Craver 2011), while others that some computational models offer fully adequate explanations without being mechanistic (e.g., Weiskopf 2011; Serban 2015; Woodward 2017).[9] One of the latest contributions to this debate is Piccinini's (2020) *Neurocognitive Mechanisms*, where he argues that all computational explanation is mechanistic (2020, 155-6) and all mechanistic explanation is constitutive explanation (2020, Sect. 6.6 and 7.1); and so, that all computational explanation is constitutive explanation. If this way of understanding Piccinini's (2020) suggestions is correct, then computational models, to the extent they are genuinely explanatory, must always uncover parts and activities constituting a target phenomenon.[10]

But, as I am about to argue, it is false that all computational explanations must be constitutive, because some fully adequate computational explanation in psychiatry (and probably in other fields, too) is aetiological and does not need to uncover constitutive mechanistic components for it to be genuinely explanatory. Distinguishing between constitutive modes of computational explanation aimed at uncovering mechanistic structure "internal to" or underlying an *explanandum* phenomenon and aetiological modes of computational explanation aimed at uncovering causal pathways and networks "extrinsic to" or preceding the phenomenon to be explained does more justice to relevant scientific practices—at least in the field of computational psychiatry—than maintaining that all adequate computational explanation is constitutive.

The distinction between aetiological and constitutive explanation is introduced by Salmon (1984), to which contemporary mechanistic philosophers generally refer (e.g., Craver 2007, 8ff; Piccinini 2020, Sect. 7.1). In developing his own account of causal explanation, Salmon (1984) says: "[i]f we want to show why [event] *E* occurred, we fill in the causally relevant processes and interactions that occupy the past light cone of *E*. This is the etiological aspect of our explanation… If we want to show why *E* manifests certain characteristics, we place inside the volume occupied by *E* the internal causal mechanisms that account for *E*'s nature. This is the constitutive aspect of our explanation" (Salmon 1984, 275).

[9] Some authors argue that there are non-causal modes of computational explanation, too. Huneman (2010) argues that there are non-causal topological explanations in biology, and Chirimuuta (2018) that some computational models in computational neuroscience are explanatory in virtue of appealing to the efficiency of the neural code, which would be a non-causal fact bearing on why certain neural circuits should possess certain information-theoretic properties. While topological and information-theoretic approaches are becoming more and more widespread in computational psychiatry, I will leave non-causal modes of computational explanation on the side in this paper.

[10] Piccinini (2020) does not explicitly say that "all mechanistic explanation is constitutive explanation" and that "all computational explanation is mechanistic," since he does not generally use quantified sentences in his account of computational explanation. For instance, he makes the non-quantified claim that "constitutive explanation, including computational explanation, is mechanistic" (156). However, because Piccinini (2020) concentrates on computational explanation as a kind of constitutive explanation, one may interpret the claim that computational explanation is mechanistic as the restricted claim that all constitutive computational explanation is mechanistic. If one also conceives of functional analyses as constitutive, then this claim would rule out that some computational explanations are not mechanistic because they are functional analyses and functional analyses are distinct and autonomous from mechanism. As it will be clear in a moment, my focus here is on whether some computational explanations are aetiological. And so, if the scope of Piccinini's account is restricted to constitutive computational explanation, my contribution here will complement Piccinini's account. Thanks to an anonymous referee for pressing me on this point.

Contemporary mechanists like Piccinini (2020) agree that aetiological and constitutive explanations are two aspects of causal explanation. In particular, aetiological causal explanations are backward-looking, as they are aimed at providing us with information about a temporal sequence of events and factors, or a pathway, leading up to an event or event type (see Ross 2021 on how pathways explanation differs from mechanistic explanation). Aetiological explanations answer questions about the occurrence of events (or event types), questions like “Why did Alf relapse?”. An answer to this question can provide us with information about some portions of the causal history of Alf, specifying some of “the factors, external to the phenomenon itself [i.e., Alf’s relapse], which bring it about” (Piccinini 2020, 156); for example, an aetiological explanation of why Alf relapsed might blame, say, his attitudes towards alcohol at a certain time, his sleeping patterns over various days prior to the relapse event and his protracted exposure to certain environmental cues. This explanation reveals an antecedent sequence of causal factors, by which Alf’s relapse occurred. This explanation can be used to devise or optimizing treatments to diminish the risk that Alf will relapse again.

Constitutive causal explanations are inward-looking, as they are aimed at providing us with information about the components and activities constituting a system’s doing something or a system’s capacity to do something. Constitutive explanations answer questions about the parts and activities constituting certain capacities or dispositions, questions like “In virtue of which constituent parts and activities is Alf disposed to relapse?”. An answer to this question provides us with information about the “internal causal mechanism” in virtue of which, say, a personal like Alf has the disposition to relapse, specifying “the internal factors that bring the phenomenon about from the inside out.” (Ibid.); for example a constitutive explanation of Alf’s disposition to relapse might refer to a certain underlying pattern of neural connectivity or atrophy in certain components in the cortico-striatal mechanism in the brain. This explanation reveals the underlying causal structure, by which a phenomenon like relapse works the way it does.

While the distinction Piccinini (2020) draws between factors “intrinsic” vs. “external” to a phenomenon is blurry, it is sufficiently clear that aetiological and constitutive explanations are both species of causal explanation and that they differ at least in terms of the kind of questions they answer and their *explananda* (Ylikoski 2013; Kaiser & Krickel 2017). It should also be obvious that knowing the causal history leading to an outcome can help researchers to identify the constitutive mechanistic components underlying the capacity responsible for the outcome, and vice versa. But this does not mean, however, that these aetiological and constitutive explanation are inseparable in practice, or that one of these two modes is more fundamental.

With this distinction between aetiological and constitutive modes of causal explanation in place, we can now define *constitutive computational explanation* as an explanation of a target phenomenon to the effect that the phenomenon is constitutively explained by at least some of the computational component and processes underlying the phenomenon. And we can define *aetiological computational explanation* as an explanation of a target phenomenon to the effect that the phenomenon is caused by at least some antecedent computational factor or process; for example, if you ask me “Why did that explosion occur at time t?”, my answer may refer to the state of a computer for generating flames at some time t −3 and I may add that the computer generated a flame at t −2, the flame ignited a firework at t − 1 and the firework exploded at t. This would be an example of an aetiological mode of computational explanation, which refers to one computational state in the past light cone of the event we want to explain.[11]

Having made some logical space for aetiological modes of computational explanations, the next question is how to characterise the sorts of explanations that RL computational modelling in psychiatry delivers. My claim is that RL models sometimes contribute to constitutive explanations of phenomena such as patients' (in)capacity to make certain choices in response to certain stimuli; but, RL models sometimes contribute also to aetiological explanations of events such as the outcome of a treatment for a certain patient.

The modelling results in Sebold et al.'s (2017) study contribute to the following explanation of why a certain patient, say Alf, relapsed vs. remained abstinent at time t: Alf relapsed because, given his expectation about how much he would enjoy an alcoholic beverage at a time t-1, Alf's *balance between model-based and model-free control* had a low value at t-1, and the interaction of these two factors led to Alf's relapse at t. Because the *explanandum* here is an event, viz. the treatment outcome for a certain patient, and because the two factors cited in the *explanans* "occupy the past light cone of" that event, the explanation on offer counts as aetiological. Because the explanation refers to a parameter defined in terms of a specific RL computational model, where different values of this parameter determine different computational states and transitions between states, the aetiological explanation sketched above is also computational.

More specifically, Sebold et al.'s (2017) evidence for the simultaneous effect of people's expectations and a specific computational parameter on the variable "risk of relapse" can tell researchers why specific individuals relapsed and why certain individuals remained abstinent. Their findings can motivate the general causal hypothesis that some alcohol-dependent patients are at higher risk of relapse, because they underestimate the reinforcing effect of drinking alcohol on their instrumental responding to alcohol-related cues, which is captured by a specific parameter in a specific RL model of decision making. In this way, RL computational modelling is used to uncover clinically relevant events and computational features that are part of the causal sequence leading up to the occurrence of an event of clinical importance like somebody's relapse. In analogy with the simple example of an aetiological computational explanation mentioned above, this would be an aetiological computational explanation, because it involves at least one computational factor that occupies the past light cone of an event to be explained.

Although one may object that Sebold et al.'s (2017) results really uncover differences between the capacities of different types of agents to remain abstinent in

[11] An anonymous referee suggested this other example of an aetiological computational explanation to me: Why did the computer output 12 at time t? Because right before t the computer calculated 5+7, the computer works properly, and 5+7=12. This referee points out that Piccinini (2020) sometimes offers examples of this sort, but Piccinini and other new mechanists interested in computational explanation do not generally notice that these sorts of examples may be best thought of as aetiological rather than constitutive explanations.

terms of different aspects of their neurocognitive endowment, this objection trades on a false dichotomy between different modes of causal explanation, and confuses the *explanandum* “Why did this patient relapse?” with the *explanandum* “In virtue of what does a patient have the capacity to remain abstinent?”. Because the use of RL modelling I have emphasised here is not aimed at the mapping of mechanisms, or at decomposing and localizing the internal causal structure of a capacity or disposition, but is aimed at answering why certain patients relapsed into alcohol abuse, RL modelling here contributes most obviously to aetiological explanation.

Of course, this is not to say that RL modelling is not useful also for identifying constitutive components of the mechanisms of relapse and of AUDs more generally. RL models represent learning and decision making processes in a given task, which might be neurally realized and can obviously explain one’s disposition to relapse. Because different components of RL models are individuated functionally—in terms of what they do in relation to a system’s capacities in a given task—an account of their organized interaction constitutes a functional explanation of psychological phenomena that abstracts away from, is noncommittal to, but for some purposes is helpfully constrained by, the way the algorithmic components, biases, mental representations and processes ascribed to patients, are physically realized in neurobiological or bodily mechanisms.[12]

In particular, Sebold et al. (2017) found that reduced activation in experimental participants’ medial prefrontal cortex (mPFC) correlated with reduced model-based control for relapsers compared to abstainers and healthy controls. This is consistent with the mechanistic hypothesis that chronic alcohol consumption causes a dysfunction of dopaminergic neurotransmission in the mesolimbic-mesocortical circuit (or “aberrant reward processing”), which would constitutively explain reduced self-control, hypersensitivity to alcohol-related cues and a susceptibility to relapse (Heinz et al. 2009). While the reinforcing effects of alcohol consumption are associated with changes in the amount of dopamine in the basal ganglia, enhanced cue-induced activity in the mPFC has also been found to predict subsequent relapse in alcohol-dependent patients (Beck et al. 2012). This type of constitutive, mechanistic information is also relevant to explain how existing pharmacological therapies work and how they might be improved (Oberlin et al. 2020).

Now, if RL modelling can contribute to both constitutive and aetiological explanations of clinically relevant phenomena, should a “full” causal explanation always include a mechanism? Piccinini (2020, 177) answers affirmatively and says that “the full mechanism at the relevant level of organization is what gives us the deepest understanding—a full explanation. It answers the largest number of what-if-things-

[12] We should be careful here to distinguish between having a computational model whose components are individuated by their neural properties, and using a model selection procedure that draws on neural evidence. Sebold et al. (2017) used brain imaging data to evaluate the degree of support of each of the algorithmic models they considered for their task, relative to each of the other models. In this way, Sebold and collaborators’ work complies with the normative suggestion that “psychologists [and psychiatrists] ought to let knowledge of neural mechanisms constrain their hypotheses just like neuroscientists ought to let knowledge of psychological functions constrain theirs” (Piccinini and Craver 2011, 285). These researchers let “knowledge of neural mechanisms” constrain the *evidential* relationships between participants’ behaviour in a given task and alternative computational models of that task. They used neural data to select the “best computational model” among a given set of models of their task.

had-been-different questions. It identifies all the difference makers at that level. It gives us the most control over the phenomenon, including how to take the system apart, how to build another like it, and how to fix it if it's broken. It's what the search for constitutive explanation ideally strives for."

Piccinini's answer here suggests, firstly, that both constitutive and aetiological explanations may be compared on the basis of their explanatory depth, which can plausibly be measured in terms of the number of what-if-things-had-been-different questions they can each answer in relation to some phenomenon of interest, and, secondly, that a causal explanation that does not aim to uncover any constitutive component of a phenomenon is defective or partial, as it would be the "full mechanism that gives us a full explanation."

My reply to the first suggestion is that it is confusing to say that mechanistic constitutive explanations are deeper or less deep than aetiological explanations, because these two species of causal explanation have different target *explananda*.[13] More importantly, my reply to the second suggestion is that some aetiological explanations provide all causal understanding one needs given certain aims. A "full" causal explanation of why a particular individual relapsed need not refer to constitutive mechanistic details, as this would yield no deeper understanding and may cost in terms of tractability and ease of understanding given the aims of improving treatment outcomes and optimizing a therapy like alcohol-avoidance training for that patient.

The choice of aims determines what "full" and "partial" understanding consists of, "how much" understanding is needed, what norms of assessment apply to a given model, and how a model should be revised or replaced due to its lack in explanatory power. Thus, being clear on why, for what purposes, one uses a certain computational model matters for understanding how one should model a target, and whether a given model is best suited to providing aetiological or mechanistic explanations (both, or neither).

Perhaps *the* central aim of psychiatry is effective mental health treatment. Some computational models produce causal insight that furthers this aim, as the insight they yield is relevant for developing, evaluating and improving, say, cognitive-behavioural therapies like alcohol-avoidance training, and for tailoring them to individual patients. One common strategy to pursue this aim is exemplified by Sebold et al.'s (2017) study. They examined patients with the same diagnosis and healthy controls, searching among them for common factors in the causal history leading to their relapse rather than abstinence. As we saw, their results lend support to the causal hypothesis that "therapeutic interventions that aim to increase goal-directed control (such as motivational interviewing) and alter the anticipated outcomes of alcohol use" probably decrease risk of relapse in particular patients (Sebold et al. 2017, 854).

[13] Piccinini (2020) does not explicitly appeal to any specific account of explanatory depth, though he probably has in mind some interventionist view, where depth is defined in terms of the range of counterfactual questions an explanatory generalization answers. This general view is not uncontentious and can be spelled out in different ways (e.g., Weslake 2010, Strevens 2011; Woodward 2021). Even assuming one can meaningfully compare the explanatory power of constitutive vs. aetiological explanations, it is actually unclear—or at least Piccinini does not provide any reason to believe—that constitutive causal explanations are always deeper than non-constitutive explanations in the context of pertinent scientific practices.

Given this sort of therapeutic aim, it is often unnecessary to map the posits of a RL model on microscopic or macroscopic properties of the brain, which would be constitutive of certain learning and decision making capacities in a given task. That's unnecessary because it would contribute no deeper understanding of why a particular patient relapsed at a certain time and place, or how alcohol-avoidance training should be optimized for decreasing the risk of relapse in a patient.

If mechanists like Piccinini (2020) maintain that all computational explanation is mechanistic; and, because all mechanistic explanation is constitutive, all computational explanation is constitutive, then they are offering an overly restrictive account of computational explanation. As there are aetiological modes of computational explanation in psychiatry, not all computational explanations are constitutive. Some computational explanations refer to computational states or processes "that occupy the past light cone" of events of interest like patients' treatment outcomes. If mechanists like Piccinini (2020) restrict their focus on constitutive explanation and acknowledge that there are fully adequate non-constitutive, aetiological causal explanations in the sciences of mind and brain, then introducing the idea of an *aetiological mode of computational explanation* would provide them with a valuable extension of their mechanistic accounts of computational explanation, as that idea helps us to better describe and understand some relevant practices at least in computational psychiatry.

5 Conclusions

In this paper, I have examined RL computational practice in psychiatry in the context of alcohol use disorders. Focusing on nosology and psychiatric explanation, I have attempted to make two claims plausible. The first is that the impact RL modelling can plausibly have on nosology is best appreciated from the vantage point of Hasok Chang's (2004) account of epistemic iteration. The second claim is that it makes sense to distinguish between two modes of computational explanation, constitutive and aetiological. While psychiatrists are often interested in both aetiological and constitutive questions, there is a diversity of aims in psychiatry. For some such aims, constitutive and aetiological modes of computational explanation need not converge on mechanism. The right epistemology to think about the roles of computational modelling in psychiatry is in terms of epistemic iteration and pluralism about computational explanation.

Acknowledgements I am grateful to Ke Chen, Andreas Heinz, Miriam Sebold and two anonymous reviewers for useful discussion and their generous comments on previous versions of this paper. This work was supported by the Alexander von Humboldt Foundation through a Humboldt Research Fellowship for Experienced Researchers at the Department of Psychiatry and Psychotherapy, at the Charité University Clinic in Berlin.

References

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition. Arlington, VA: American Psychiatric Association

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition. Washington, DC: American Psychiatric Association

American Psychiatric Association (1980). *Diagnostic and Statistical Manual of Mental Disorders*, Third Edition. Washington, DC: American Psychiatric Association

American Psychiatric Association (1968). *Diagnostic and Statistical Manual of Mental Disorders*, Second Edition. Washington, DC: American Psychiatric Association

American Psychiatric Association (1952). *Diagnostic and Statistical Manual of Mental Disorders*, First Edition. Washington, DC: American Psychiatric Association

Batra, A., Müller, C. A., Mann, K., & Heinz, A. (2016). Alcohol Dependence and Harmful Use of Alcohol: Diagnosis and Treatment Options. *Deutsches Ärzteblatt International*, 113(17), 301–310

Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: A neurocognitive perspective. *Nature Neuroscience*, 8, 1458–1463

Beck, A., Wüstenberg, T., Genauck, A., Wrase, J., Schlagenhauf, F., Smolka, M. N. … Heinz, A. (2012). Effect of brain structure, brain function, and brain connectivity on relapse in alcohol-dependent patients. *Archives of general psychiatry*, 69(8), 842–852

Boffo, M., Zerhouni, O., Gronau, Q. F., van Beek, R. J., Nikolaou, K., Marsman, M., & Wiers, R. W. (2019). Cognitive bias modification for behavior change in alcohol and smoking addiction: Bayesian meta-analysis of individual participant data. *Neuropsychology review*, 29(1), 52–78

Brandon, T. H., Vidrine, J. I., & Litvin, E. B. (2007). Relapse and relapse prevention. *Annual Review of Clinical Psychology*, 3, 257–284

Brown, V. M., Chen, J., Gillan, C. M., & Price, R. B. (2020). Improving the reliability of computational analyses: Model-based planning and its relationship with compulsivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*

Carvalho, A. F., Heilig, M., Perez, A., Probst, C., & Rehm, J. (2019). Alcohol use disorders. *The Lancet*, 394(10200), 781–792

Chang, H. (2017). Epistemic iteration and natural kinds: Realism and pluralism in taxonomy. In K. Kendler, & J. Parnas (Eds.), *Issues in psychiatry IV: Classification of psychiatric illnesses* (pp. 229–245). Oxford: Oxford University Press

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press

Chirimuuta, M. (2018). Explanation in computational neuroscience: Causal and non-causal. *The British Journal for the Philosophy of Science*, 69(3), 849–880

Colombo, M. (2021). (Mis)computation in Computational Psychiatry. In F. Calzavarini & M. Viola (Eds.). *Neural Mechanisms. New Challenges in the Philosophy of Neuroscience* (pp. 427–448). Dordrecht: Springer Studies in Brain and Mind 17

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

Colombo, M., & Heinz, A. (2019). Explanatory integration, computational phenotypes, and dimensional psychiatry: The case of alcohol use disorder. *Theory & Psychology*, 29(5), 697–718

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: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press

Craver, C., & Tabery, J. (2015). Mechanisms in Science. In *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL = < https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/>

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., & Barch, D. M. (2016). Reduced model-based decision-making in schizophrenia. *Journal of Abnormal Psychology*, 125, 777–787

Cuthbert, B. N. (2014). The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, 13(1), 28–35

Dalgleish, T., Black, M., Johnston, D., & Bevan, A. (2020). Transdiagnostic approaches to mental health problems: Current status and future directions. *Journal of Consulting and Clinical Psychology*, 88(3), 179–195

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(6), 1204–1215

Dayan, P., & Balleine, B. (2002). Reward, motivation, and reinforcement learning. *Neuron*, 36(2), 285–298

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, 80(2), 312–325

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, 8, 1481–1489

Flanagan, O. (2013). Identity and addiction: What alcoholic memoirs teach. In K. W. M, Fulford, et al. (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 865–888). Oxford: Oxford University Press

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

Grob, G. N. (1991). Origins of DSM-I: A study in appearance and reality. *American Journal of Psychiatry*, 148(4), 421–431

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., Beck, A., Grüsser, S. M., Grace, A. A., & Wrase, J. (2009). Identifying the neural circuitry of alcohol craving and relapse vulnerability. *Addiction biology*, 14(1), 108–118

Huneman, P. (2010). Topological explanations and robustness in biological sciences. *Synthese*, 177, 213–245

Huys, Q. J. M., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacol*, 46, 3–19

Jellinek, E. M. (1960). *The Disease Concept of Alcoholism*. New Brunswick, NJ: Hillhouse Press

Kaiser, M. I., & Krickel, B. (2017). The metaphysics of constitutive mechanistic phenomena. *The British Journal for the Philosophy of Science*, 68(3), 745–779

Kakoschke, N., Kemps, E., & Tiggemann, M. (2017). Approach bias modification training and consumption: A review of the literature. *Addictive behaviors*, *64*, 21 – 8

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183, 339–373

Kellert, S., Longino, H., & Waters, C. K. (2006). Introduction: The pluralist stance. In S.H. Kellert, H.E. Longino, C.K. Waters (Eds.), *Minnesota Studies in Philosophy of Science, vol. 19: Scientific Pluralism*, University of Minnesota Press, Minneapolis (2006), pp. vii-xxix

Kendler, K. S. (2012). The dappled nature of causes of psychiatric illness: replacing the organic- functional/ hardware- software dichotomy with empirically based pluralism. *Molecular Psychiatry*, 17, 377–388

Laurens, M. C., Pieterse, M. E., Brusse-Keizer, M., Salemink, E., Allouch, S. B., Bohlmeijer, E. T., & Postel, M. G. (2020). Alcohol avoidance training as a mobile app for problem drinkers: longitudinal feasibility study.JMIR mHealth and uHealth, 8(4), e16217

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, 14(2), 154–162

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

Moos, R. H., & Moos, B. S. (2006). Rates and predictors of relapse after natural and treated remission from alcohol use disorders. *Addiction*, 101(2), 212–222

Morris, S. E., & Cuthbert, B. N. (2012). Research Domain Criteria: cognitive systems, neural circuits, and dimensions of behavior. *Dialogues in clinical neuroscience*, 14(1), 29

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

Nathan, P. E., Conrad, M., & Skinstad, A. H. (2016). History of the Concept of Addiction. *Annual review of clinical psychology*, 12, 29–51

Oberlin, B. G., Shen, Y. I., & Kareken, D. A. (2020). Alcohol Use Disorder Interventions Targeting Brain Sites for Both Conditioned Reward and Delayed Gratification. *Neurotherapeutics*, 17(1), 70–86

Patzelt, E., Hartley, C., & Gershman, S. (2018). Computational Phenotyping: Using Models to Understand Individual Differences in Personality, Development, and Mental Illness. *Personality Neuroscience*, 1, E18. doi:https://doi.org/10.1017/pen.2018.14

Piccinini. (2020). *Neurocognitive Mechanisms*. Oxford University Press

Piccinini, G., & Craver, C. (2011). Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches. *Synthese*, 183(3), 283–311

Rehm, J., & Room, R. (2015). Cultural specificity in alcohol use disorders.Lancet, S0140–6736

Rehm, J., Marmet, S., Anderson, P., Gual, A., Kraus, L., Nutt, D. J. … Gmel, G. (2013). Defining substance use disorders: do we really need more than heavy use? *Alcohol and alcoholism*, 48(6), 633–640

Rehm, J., Baliunas, D., Borges, G. L., Graham, K., Irving, H., Kehoe, T. … Taylor, B. (2010). The relation between different dimensions of alcohol consumption and burden of disease: an overview. *Addiction*, 105(5), 817–843

Robinson, S. M., & Adinoff, B. (2016). The classification of substance use disorders: Historical, contextual, and conceptual considerations. *Behavioral Sciences*, 6(3), 18

Room, R. (2005). Stigma, social inequality and alcohol and drug use. *Drug and alcohol review*, 24(2), 143–155

Ross, L. N. (2021). Causal concepts in biology: How pathways differ from mechanisms and why it matters. *The British Journal for the Philosophy of Science*, 72(1), 131–158

Rush, B. (1784/1811). *An Inquiry into the Effects of Ardent Spirits upon the Human Body and Mind, with an account of the means of preventing, and of the remedies for curing them. Sixth edition*. New-York: Printed for Cornelius Davis

Saha, T., Chou, P., & Grant, B. (2006). Toward an alcohol use disorder continuum using item response theory: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Psychological Medicine*, 36, 931–941

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University

Schomerus, G., Lucht, M., Holzinger, A., Matschinger, H., Carta, M. G., & Angermeyer, M. C. (2011). The stigma of alcohol dependence compared with other mental disorders: a review of population studies. *Alcohol and alcoholism*, 46(2), 105–112

Seriés, P. (Ed.). (2020). *Computational psychiatry: A primer*. MIT Press

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

Sellman, J. D., Foulds, J. A., Adamson, S. J., Todd, F. C., & Deering, D. E. (2014). DSM-5 alcoholism: a 60-year perspective. *Australian & New Zealand Journal of Psychiatry*, 48(6), 507–511

Serban, M. (2015). The scope and limits of a mechanistic view of computational explanation. *Synthese*, 192(10), 3371–3396

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: rationale and reliability. *Archives of general psychiatry*, 35(6), 773–782

Strevens, M. (2011). *Depth: An account of scientific explanation*. Harvard University Press

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press

Tsou, J. Y. (2012). Intervention, causal reasoning, and the neurobiology of mental disorders: Pharmacological drugs as experimental instruments. *Studies in History and Philosophy of Science Part C*, 43(2), 542–551

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. … Bullmore, E. T. (2015). Disorders of compulsivity: A common bias towards learning habits. *Molecular Psychiatry*, 20, 345–352

Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese*, 183, 313–338

Weslake, B. (2010). Explanatory depth. *Philosophy of Science*, 77(2), 273–294

Wiers, R. W., Van Dessel, P., & Köpetz, C. (2020). ABC Training: A New Theory-Based Form of Cognitive-Bias Modification to Foster Automatization of Alternative Choices in the Treatment of Addiction and Related Disorders. *Current Directions in Psychological Science*, 29(5), 499–505

Wiers, R. W., Boffo, M., & Field, M. (2018). What's in a trial? On the importance of distinguishing between experimental lab studies and randomized controlled trials: the case of cognitive bias modification and alcohol use disorders. *Journal of Studies on Alcohol and Drugs*, 79(3), 333–343

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

Witkiewitz, K., Litten, R. Z., & Leggio, L. (2019). Advances in the science and treatment of alcohol use disorder. *Science Advances*, 5(9), eaax4043

Woodward, J. (2021). Explanatory autonomy: the role of proportionality, stability, and conditional irrelevance. *Synthese*, 198, 237–265

Woodward, J. F. (2003). *Making Things Happen*. New York: Oxford University Press

World Health Organization (2020). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*, 11th revision. Geneva: World Health Organization

World Health Organization. (2018). *Global Status Report on Alcohol and Health*. Geneva: WHO Press

Ylikoski, P. (2013). Causal and constitutive explanation compared. *Erkenntnis*, 78(2), 277–297