(*Mis*)*computation* in Computational Psychiatry

Matteo Colombo
https://mteocolphi.wordpress.com/
m.colombo @ uvt .nl

Abstract An adequate explication of *miscomputation* should do justice to the practices involved in the computational sciences. As relevant practices outside computer science have been overlooked, I begin to fill this gap by distinguishing different notions of *miscomputation* in computational psychiatry. I argue that a satisfactory explication of *miscomputation* in computational psychiatry should be grounded in the semantic view of computation, rather than in the mechanistic view. To the extent my argument is convincing, we should reconsider the adequacy of the mechanistic view of computation for illuminating some methodological and explanatory practices in computational cognitive neuroscience, as well as for individuating biological computing systems.

Keywords: miscomputation; computational psychiatry; mechanism; semantics

1. Introduction

The concept *computer* was originally used to mean "individual people carrying out calculations"—for example, to refer to people performing calculations of the calendar in medieval Europe (Uckelman 2018), or to hierarchically organized factories of human workers manufacturing logarithmic and trigonometric tables in France in the 1790s (Daston 1994). At least since Descartes, the notion of *computation* has increasingly been employed for understanding how thinking works and how it is related to physical processes (Isaac 2019). Contemporary understanding of *computation*, and of its relevance for understanding how thinking works, stems in large part from Alan Turing's (1936, 1950) work on whether or not there is some mechanical procedure (i.e., an algorithm) that can correctly decide the truth value of any arbitrary mathematical formula.

In any of its uses throughout history, c*omputation* refers to a kind of rule-governed operation. This operation consists in transitions from inputs to outputs, following a definite rule defined over formal properties of the possible inputs and outputs of a system. Because computing systems would thus be kinds of rule-governed systems, they can make errors, that is, their operations can violate the rules they are supposed to follow.

An adequate explication[1] of the concept of *miscomputation* should make more precise the generic, pre-scientific explicandum concept of *error in computing* and fruitfully illuminate existing

[1] According to Carnap (1950, 3), the task of explication consists in "transforming a given more or less inexact concept into an exact one, or, rather, in replacing the first by the second." Carnap (1950, 7) holds that an adequate explicatum should be similar to the explicandum in respecting prior usage—though "close similarity is not required" and "considerable differences are permitted." It should be more exact than the explicandum. It should be fruitful in the sense of being "useful for the formulation of […] empirical laws [or] logical theorems." And last, the explicatum should be simple. One example Carnap (1950, 5-6) offers is the explication of the

scientific practices, where this concept is employed for various epistemic and practical purposes. An explicatum may be specified by giving an explicit definition, or by providing rules for apt usage. Below I shall specify two explicata of *miscomputation* by giving two definitions.

While the literature in the philosophy of computation has been concerned with the ontological task of individuating physical computing systems, with providing identity conditions for concrete computing systems (e.g., Chalmers 2011; Miłkowski 2013; Fresco 2014; Piccinini 2015; Coelho Mollo 2018; Lee 2018; Shagrir 2018; Schweizer 2019), in this paper I focus on the concept of *error in computing*, or *miscomputation*. I distinguish different uses of this concept in computational psychiatry, and argue that at least one satisfactory explication of *miscomputation* requires an appeal to semantic properties, which fix the identity of the rule-governed operations assumed to be carried out by brain mechanisms of interest.

I want to show that some ascriptions of *miscomputation* in contemporary computational psychiatry should be understood as semantically-laden, interest-relative and perspectival, although non-arbitrary, relatively clear-cut, experimentally evaluable, and instrumentally useful. If any concept of *computation* entails the concept of *miscomputation*, and at least one adequate explication of *miscomputation* should always refer to some semantic properties of a system of interest, then at least one adequate explication of *computation* cannot do without semantics either.

I begin by outlining the main aims and approaches of contemporary computational psychiatry, distinguishing four contexts where some normative concept is employed to characterise some mental illnesses (Section 2). Then, I lay out two possible explications of *miscomputation* grounded in the semantic view and the mechanistic view of computation (Section 3). On the basis of a representative case study, I argue that at least one satisfactory explication of *miscomputation* in computational psychiatry should involve a semantically-laden characterisation of biological computing systems' interaction with their environment (Section 4). I conclude by summarizing my argument, and outlining two consequences for the mechanistic view of computation (Conclusion).

2. *Miscomputation* in computational psychiatry

Contemporary psychiatry faces many explanatory gaps. Classification and diagnosis are unreliable, unspecific and offer little guidance about treatment (Stephan et al. 2016a). The physiological, psychological and social mechanisms of mental illnesses remain ill-understood, while psychiatrists

---

prescientific concept of *fish*, understood as an animal living in water, in terms of the scientific explicatum: *animal which lives in water, is a cold-blooded vertebrate, and has gills throughout life*. Kemeny and Oppenheim (1952, 308), on the other hand, distinguished their project from Carnap's in these terms: "The commonest procedure of explication is to apply a trial and error method till one arrives at an ingenious guess, and then try to find intuitive reasons to justify the proposed explicatum. This procedure is clearly very dangerous: The intuition of the most honest and well-trained philosopher is likely at times to become a tool for grinding an axe. […] We feel that we must first put down clearly all that our intuition tells us about the explicandum, and then find the precise definitions that satisfy our intuitive requirements."

and philosophers of psychiatry tend to unjustifiably conceive of their aetiologies in dualistic terms, as either brain-based or mind-based (Stephan et al. 2016b). These (and other) factors prevent cross-validation of mental illnesses among genetic markers, brain mechanisms and clinical assessment, creating an explanatory gap between theoretical models and causal knowledge of mental illnesses, and the categories and treatments used in clinical practice.

In an attempt to bridge this gap, computational approaches to psychiatry have blossomed in the last few years. Grounded in the ideas that mental states can be modelled as computational states and that computation can explain mental phenomena, computational psychiatry aims to identify the computational "aberrations" responsible for specific clusters of psychiatric symptoms (Stephan et al. 2016a, 79; Fletcher & Frith 2009; Brugger & Broome 2019). To pursue this aim, computational psychiatrists have started to identify *computational phenotypes* (i.e., measurable behavioural and neural types defined in terms of specific parameters extracted from specific computational models of a given task) associated with mental illness that straddle conventional diagnostic classifications (Montague, Dolan, Friston & Dayan 2012; Patzelt, Hartley & Gershman 2018; Colombo & Heinz forthcoming).

Computational psychiatry assumes that the neural and behavioural data researchers collect during experiments with psychiatric patients and/or healthy controls are generated by one or several computational processes. To identify these processes, computational psychiatrists rely on data-driven or theory-driven methodologies. These methodologies allow computational psychiatrists to mine large sets of data for identifying structures that might help them with reliable diagnostic classification, or to simulate experimental participants' behaviour and brain activity in an attempt to identify promising ways for treating mental illnesses (Huys, Maya & Frank 2016).

Typical in data-driven methodologies is the use of machine-learning techniques to mine large sets of neural and behavioural data from psychiatric patients and healthy controls, for patterns, clusters, and causal dependencies that may aid diagnostic classification of illnesses and prediction of treatment outcomes. Data-driven methodologies do not generally involve hypotheses about computational processes and mechanisms.

Theory-driven methodologies use computational models embodying prior knowledge of, or hypotheses about, neurocomputational mechanisms. Computational models are used to precisely relate existing clinical concepts like *schizophrenia*, *delusion*, *autism* and *anxiety* to specific parameters of the models, which can in turn be related to measurable aspects of behaviour and neural activity. The purpose is to seek "to characterize mental dysfunction in terms of aberrant computations over multiple scales" (Montague et al. 2012, 72).

Three broad classes of approaches can be distinguished in contemporary computational psychiatry. One class relies on both data-driven and theory-driven methodologies from network

science (Sporns 2010, Ch. 10). The aim is to identify topological, anatomical and causal properties of brain networks, which could provide psychiatrists with biomarkers for diagnosis and for the evaluation of treatment efficacy. In the context of network approaches, one guiding hypothesis is that specific patterns of "'faulty wiring' or aberrant connectivity in the brains" are responsible for mental illnesses, which are conceived of as dysconnectivity syndromes (Cao, Wang & He 2015, 2801).

Two classes of approaches that typically rely on theory-driven methodologies include biophysically detailed connectionist modelling, and modelling based on Bayesian decision theory and Reinforcement Learning (RL) (Sutton & Barto 1998). Biophysically detailed connectionist models aim to explain mental illnesses in terms of the emergent behaviour of dynamically interacting neuron-like units, each one of which computes simple mathematical functions that refer to physical properties of the units like their voltage and conductance. Connectionist models have been used to examine how changes in neural connectivity and levels of neuromodulatory transmitters affect network dynamics, and to predict patients' performance in a battery of behavioural and cognitive tasks (e.g., Cohen & Servan-Schreiber 1992) or patterns of neural activity observed in mental conditions like epilepsy (e.g., Ursino & La Cara 2006). In the context of connectionist approaches, one guiding hypothesis is that aberrant network architecture and electrical dynamics within networks of neurons are responsible for mental illnesses, which are conceived of as emergent "disordered" dynamics within local and distributed neural networks (Globus & Arpaia 1994).

Models from RL and Bayesian decision theory provide computational psychiatrists with trial-by-trial hypotheses about how a given decision-making or learning task may be solved. These hypotheses consist in algorithmic models, which describe the operations participants might carry out to solve a given experimental task. By fitting algorithmic models to choice and neural data, and evaluating their relative degree of support, computational psychiatrists extract values of model parameters, which they can use to characterise the processes producing participants' choices as "(sub)optimal," "statistically abnormal" or "aberrant," to cluster and classify different performance trajectories of the participants in the task, and to correlate observed behaviour, neural activity, known psychiatric symptoms and hypothesized computations (Montague et al. 2012; Patzelt, Hartley & Gershman 2018). In the context of Bayesian and RL approaches, one guiding hypothesis is that mental illnesses, which are conceived of as aberrant computational phenotypes (Colombo & Heinz forthcoming), are due to systematic misrepresentations of one's (social) environment or bodily state, or to computational processes, where "mechanisms of *inference and integration* might be suboptimal or deviant, leading to incorrect estimates of states or choices of action" (Huys, Guitart-Masip, Dolan & Dayan 2015, 401).

This overview highlights that there are at least four contexts in computational psychiatric practice, where some normative concept is used to characterise mental illnesses. These normative notions include:

i. dysfunctional brain connectivity
ii. "disturbed" network dynamics
iii. misrepresentation
iv. "aberrant" processes of inference, information integration, or choice

Among these four notions, only the fourth and the second are obviously associated with an error in computing. After all, the modelling contexts where they feature generally assume that brains (or parts therefor) carry out rule-governed operations. In what follows, I focus on (iv), and ask whether either the mechanistic or the semantic view of computation can provide us with the conceptual resources for adequately explicating *miscomputation* as understood in (iv).

3. Explicating *miscomputation*. Mechanistic and semantic style

The mechanistic and semantic views of computation concern the identity conditions of computing systems. These two views can be useful also for helping us explicate the concept of *miscomputation*. In fact, it has been claimed that "[w]hile it is hard or impossible to make sense of miscomputation within traditional accounts of computation [like semantic accounts], miscomputation finds an adequate explication within the mechanistic account", thereby allowing us to adequately account for the language and practices of the computational sciences (Piccinini 2015, 275).

3.1 The mechanistic view of *miscomputation*

Based on the mechanistic view, the explicandum concept *miscomputation*, i.e., an error in computation, can be explicated as:

miscomputation^{M}: A computing system miscomputes just in case the system malfunctions, where what counts as a malfunction is not fixed by any semantic property, but by the objective goal or selective history of the system, which fixes the teleological functions to compute possessed by systems of that type.

Let's clarify how this explicatum is grounded in the conceptual resources of the mechanistic view. According to this view, computing systems are species of mechanisms, that is: spatially and temporally organized structures, performing a function in virtue of their component parts, causal activities and organization (Bechtel 2008, 13).

One function of computing mechanisms is that of performing computations (e.g., Miłkowski 2013; Piccinini 2015; Coelho Mollo 2018). Specifically, computing systems possess at least one teleological function to compute. The teleological functions of a type of mechanism are determined by the causal contributions of some of the dispositions and features of mechanisms of that type to an objective goal (Maley & Piccinini 2017) or to the present existence and persistence through history of mechanisms of that type (Neander 1991). For example, the teleological function(s) of the brain of an organism in a population would be a stable causal contribution (whatever that is) of that brain to that organism's objective goals of survival, reproduction and inclusive fitness, or to the present existence and persistence through history of organisms in that population.

For a mechanism like the heart, whose functioning is better understood than that of a brain, its teleological function is now uncontroversial and consists in pumping blood at a certain rate.[2] This ascription can be justified in one of two ways, depending on whether one gives more emphasis to the objective goal or to the evolutionary history of organisms with hearts. One may say that the teleological function of hearts is to pump blood, because pumping blood makes a stable causal contribution to organisms' objective goals of survival, reproduction and inclusive fitness, or because organisms with hearts pumping blood had an evolutionary history that explains the present existence, and persistence over time, of organisms with hearts pumping blood.

At least one of the teleological functions performed by computing systems must consist in the manipulation of medium-independent vehicles "based solely on differences between different portions of the vehicles according to a rule defined over the vehicles" (Piccinini 2015, 1; Coehlo Mollo forthcoming). Vehicles of computation are physical variables like neural spike trains (Maley 2018), whose values can be input to a system, can be changed by the system, or be outputs of the system. Vehicles are medium-independent, if their individuation does not depend on their physical constitution, but only on their functional profile. For example, in the case of neural spike trains, on their rate, timing or count in a given time window.

While computing systems' manipulation of medium-independent vehicles is sensitive only to specific dimensions of variation (or degrees of freedom) of the vehicles, the rules governing these manipulations are defined over the relevant degrees of freedom of relevant vehicles. While there is no consensus on how the teleological function of a computing system should be individuated (see, e.g., Dewhurst 2018; Tucker 2018; Coehlo Mollo forthcoming), a relevant teleological function of the system determines what rules ought to govern its operations over medium-independent vehicles. In the case of artefacts like desktop computers, the relevant teleological function is fixed by a

[2] Understanding that the teleological function of the heart is to pump blood was an achievement. Before William Harvey's ideas came to be widely accepted towards the end of the seventeenth century, anatomical knowledge of the heart was relatively detailed, but its actual teleological function was ill understood and different epistemic communities ascribed different objective functions to it.

specification at some level of abstraction, which is established by some computer scientist, designer or programmer with certain intentions and purposes, working in a research community with accepted standards and blueprints for the design, development and programming of desktop computers (Fresco & Primiero 2013). In the case of biological computing systems, what rules a system ought to follow is determined by the system's evolutionary history or objective goal (Piccinini 2015 Ch. 6-7).

Because malfunctions are deviations from teleological functions, and not all malfunctions are miscomputations, some account of actual functions computed by a system is required for applying *miscomputation*^{M} aptly. Like in the case of teleological functions, there's no consensus among proponents of the mechanistic view about how we should individuate what a computing system actually computes at a time. For example, Tucker (2018, 8) argues that a system's computational structure is individuated without any reference to factors external to the system; what the system is actually computing at a time is determined by the actual inputs to the system at that time in addition to its computational structure (i.e., how the system manipulates those inputs and transforms them into outputs). Piccinini argues that a computing system's teleological functional properties are individuated widely, by considering the system's objective goal, and suggests that what the system actually computes at a time is individuated "parasitically" on the system's teleological function (cf., Tucker 2018, 12).

3.2 The semantic view of *miscomputation*

Based on the semantic view of computation, the explicandum concept *miscomputation*, i.e., an error in computation, can be explicated as:

miscomputation^{S}: A computing system miscomputes just in case what the system does violates what it ought to do, where the normative and actual behaviour of the system are fixed by the semantic properties justifiably ascribed to the system's vehicles of computation.

Let's consider how this explicatum is grounded in the semantic view. The central idea of the semantic view is that some semantic properties are always necessary to individuate computing systems, and also to determine what rules actually govern their input-output transitions, and what rules ought to govern them (e.g., Fodor 1975; Churchland & Sejnowski 1992; Sprevak 2010; Rescorla 2014a; Shagrir 2018).

Like the mechanistic view, the semantic view claims that computation consists in the manipulation of medium-independent physical vehicles. Unlike the mechanistic view, however, it asserts that computing vehicles are partly individuated by what they represent, by the content they

carry, by what they stand in for. If a vehicle is not representational, then it is not computational. Thus, computing systems are distinguished from non-computing systems because computing systems can manipulate representations, while non-computing systems cannot. For example, unless a device can manipulate representations that stand in for numbers, that device cannot be a computer performing mathematical calculations.

Different computing systems operate on different kinds of representations with different formats or with different types of representational content. In particular, some representations may bear content that is perspective-dependent, or as Shagrir calls it "interpretative", that is, content whose ascription is grounded in the epistemic practices of relevant scientific communities at a given historical time (cf., Dennett 1987, Egan 2019).

The semantic content of computational vehicles can be fixed widely, by relevant relational properties of the system in a certain environment, or narrowly, by intrinsic properties of the system. If content is determined widely—for example by counterfactual causal-informational dependencies holding between a vehicle and the environment (Dretske 1981), the evolutionary function of the vehicle (Millikan 1984), or some relevant linguistic community (Kripke 1982)—physically identical computational vehicles embedded in different social or physical environments may carry different content, and hence individuate different computing systems.

However semantic content is fixed, computing systems' manipulations of representational vehicles would generally be causally sensitive only to formal properties of the vehicles (for a nuanced, qualified understanding of this claim, see Rescorla 2014b). At the same time, however, computing systems' manipulations would preserve the semantic properties of the vehicles (Fodor 1975). Semantic views of computation typically hold there's some morphism between the semantic and formal properties of vehicles, so that computing operations can preserve the intelligibility and coherence of the contents carried by the vehicles.

While within the mechanistic view what rules do govern and what rules ought to govern the transitions of a computing system are individuated functionally, on the basis of non-semantic, causal-historical properties of the computing system, computational rules are individuated semantically within the semantic view, on the basis of the content ascribed to the system's vehicles. The idea is that assumptions about the content of input and output vehicles are required to correctly determine the task performed by the system, as well as the task the system ought to perform (e.g., Sprevak 2010; Shagrir 2018). These semantically individuated tasks would fix the rules governing the system's behaviour; but, this idea doesn't commit proponents of the semantic view to claim that these rules must be explicitly represented by the computing system. Some rules can be hardwired—either because of a design plan devised by human programmers or because of biological evolution—or they can be "implicitly" derived from the contents of relevant representations manipulated by the system.

4. Explicating *miscomputation* in computational psychiatry

The question whether *miscomputation*^{M} or *miscomputation*^{S} can provide us with an adequate explication of *miscomputation* in contemporary computational psychiatry hinges on the role of semantic considerations in determining computational vehicles and in fixing the rules followed by a computing system.

Piccinini, one of the advocates of the mechanistic view, distinguishes five notions of miscomputation, namely: a mistake in computer design, a mistake in manufacturing, a mistake in programming, faulty usage, and a hardware failure (2007, 523–4; 2015, 149-50). As Fresco & Primiero (2013) and Dewhurst (2014) have noted, this list conflates semantic and non-semantic characterisations of miscomputation; but Piccinini (2015, 149-50) is explicit that only the last notion on his list—the notion of miscomputation as a "failure of a hardware component to perform its function"—is a genuine malfunction of the computing mechanism itself. This notion corresponds to *miscomputation*^{M}.

Hardware failure in computing systems is responsible for what Fresco and Primiero (2013) call "operational errors,"[3] and Turing (1950, 449) calls "errors of functioning."[4] This type of error depends on physical faults of computational or non-computational components or activities of a system, which cause it to produce an output o_2 different from the output of the function f on input i, $f(i) = o_1$, it ought to compute (Piccinini 2015, 13).

Four more specific species of hardware failure responsible for miscomputations can be distinguished. One species consists in some (non-essential) component of the computing system being missing. Another in some (non-essential) change in the structural, physical organization of the components. Another in the altered strength of the causal activities of some components. Yet another species of hardware malfunction consists in the altered temporal organization of the causal

[3] Focusing on the practices of computer science and engineering, Fresco and Primiero (2013) helpfully distinguish different levels of abstraction at which computational errors can be understood. As it will become clear, the most pertinent levels of abstraction for explicating miscomputation in computational psychiatry are what Fresco and Primiero call "the functional specification level," where a computational problem and problem domain are defined, and the levels of "algorithm implementation" and "algorithm execution", which may involve some hardware component failing to perform its designated operation because of faulty hardware design or exposure of the system to external forces like radiation, heat or friction.

[4] Writes Turing: "We may call [… these two types of errors] 'errors of functioning' and 'errors of conclusion'. Errors of functioning are due to some mechanical or electrical fault which causes the machine to behave otherwise than it was designed to do. In philosophical discussions one likes to ignore the possibility of such errors; one is therefore discussing 'abstract machines'. These abstract machines are mathematical fictions rather than physical objects. By definition they are incapable of errors of functioning. In this sense we can truly say that 'machines can never make mistakes'. Errors of conclusion can only arise when some meaning is attached to the output signals from the machine. […] When a false proposition is typed we say that the machine has committed an error of conclusion. There is clearly no reason at all for saying that a machine cannot make this kind of mistake." (Turing 1950, 449).

activities of some components. These species of failure are the focus of much of the current research in neuropsychology and neurology (Barack & Platt 2016).

If we consider neuropsychological research focused on differences in brain connectivity between mentally ill patients and healthy controls. Brain connectivity refers to patterns of anatomical links or of causal interactions between distinct units at different scales within the nervous system. Psychiatrists pursuing this approach typically relate mental illnesses to specific units, or specific anatomical connections between units, being missing or altered. For example, several mental illnesses including obsessive-compulsive disorder, mania, and major depressive disorder emerge after traumas in frontal and subcortical areas (Schwarzbold et al. 2008); central symptoms of schizophrenia are associated with structural connectivity reductions in the frontal lobe, compared with healthy controls (Petterson-Yeo et al. 2011); and brains of patients with major depressive disorder present causal relationships between orbitofrontal cortex, the temporal lobe, cortical areas, and the hippocampus that are altered relative to control subjects (Rolls et al. 2018).

However, because these results do not rely on the assumption that brains are kinds of computing systems, no notion of *miscomputation* (or *error in computing*) is typically used to characterize them. A network approach can be combined with this assumption (for examples see Stephan, Baldeweg & Friston 2006; Wang & Krystal 2014). But, where researchers make no assumption that observed behavioural and neural data are generated by computational transformations carried out by brains and rely on no computational method, we should be careful to interpret their claims about brain dysfunction in terms of the concept of *miscomputation*. For this concept can be employed most aptly in the context of practices that are grounded in the idea that the brain is a computing system, or in computational methods for probing these brain processes conceived of as computations.

One such context involves RL or Bayesian modelling. Recall this approach involves claims that some mental illnesses are due to "false", "suboptimal" or "aberrant" processes of inference, information integration, or choice. These types of claims generally inform the research question, design, and interpretation of the results of typical studies aimed at probing the computational operations psychiatric patients and healthy controls might deploy for solving a certain learning or decision-making task. Schlagenhauf et al.'s (2014) study of striatal dysfunction in schizophrenia during reversal learning tasks provides us with a representative case study, which will be helpful for evaluating the adequacy of *miscomputation*^{M} and *miscomputation*^{S}.

4.1 A case study

The question Schlagenhauf et al. (2014) asked is whether differences in brain and behavioural data can distinguish abnormal computations diagnostic of mental illness from abnormal computations

that only indicate differences in the operations to solve a given task. They addressed this question by using a model-based fMRI design (Colombo 2014b).

Specifically, they collected behavioural and neural data from patients with schizophrenia and healthy controls undergoing magnetic resonance brain imaging, while they performed a probabilistic reversal learning task. This task requires participants to learn from probabilistic feedback, where the structure of the task can change so that what used to be positive outcomes (i.e., a positive reward) are now negative outcomes (i.e., a punishment, or negative reward) and what used to be negative are now positive outcomes.

Schlagenhauf et al. (2014) formulated three types of RL algorithmic models of the task, which corresponded to different hypotheses about the computational operations underlying individual choices. Two models were variants of the Rescorla-Wagner model, where choices are made on the basis of the expected value of taking an action on a trial and of model parameters; expected actions values are updated after each trial through prediction errors signalling the difference between expected outcome and actually received outcome. The other model was a Hidden Markov Model (HMM), which builds a model of the task via (state) prediction errors; this model of the task can be used to infer the (hidden) state of the current trial; and inferred states, along with model parameters, determine choices.

Schlagenhauf and colleagues used Bayesian model selection to determine the degree of support of each of the three models, given individual participants' trial-by-trial behavioural and neural data. This procedure allowed them to identify a HMM with differential reward and punishment sensitivity as the best supported hypothesis about the processes underlying choices both in healthy controls and in several of the schizophrenia patients. For the other schizophrenia patients, the best fitting hypothesis was a Rescorla-Wagner model.

Model-based fMRI and Bayesian model comparison allowed the researchers to identify brain activity strongly associated with trial-by-trial variation in parameters of each computational model, and in particular to quantify the magnitude of brain activity signalling prediction errors in individual participants. It was found that, in response to similar patterns of state-reward contingencies, both groups of schizophrenia patients exhibited reduced prediction error signalling in the ventral striatum, compared to controls. It was also found that patients, whose computational processes were well-modelled by the HMM, showed patterns of prefrontal activity similar to those of healthy controls; but these patients' prefrontal activity was higher than that observed in patients whose computational processes were modelled by the Rescorla-Wagner model.

The finding of reduced prediction error signalling in the ventral striatum in both groups of schizophrenia patients was taken to indicate that "aberrant" prediction error signalling is responsible for at least some of the core deficits in schizophrenia. More specifically, Schlagenhauf et al. (2014)

took their results to provide evidence for two conclusions. First, dysfunction of prediction error signalling in the striatum, or "aberrant prediction error," is a core miscomputation responsible for pathological learning deficits in schizophrenia patients, even when patients use computational operations of the *same type* as those of healthy controls. Second, the same population of schizophrenia patients can use *different types* of computational operations to solve a given task, even when they present a common pattern of psychiatric symptoms.

Schlagenhauf et al.'s (2014) research question, design, and interpretation of results in terms of "aberrant prediction error" are not fruitfully illuminated by a notion of *miscomputation*^{M} as hardware malfunction; *miscomputation*^{S} can instead help us make good sense of their practices. Because *miscomputation*^{S} but not *miscomputation*^{M} can do justice to relevant practice in at least one prominent approach in computational psychiatry, we have reason to favour a semantically-laden explication of *miscomputation*. Let's now flesh out this argument.

4.2 Why (*mis*)*computation* requires semantics

How should we understand the notion of *miscomputation* used by Schlagenhauf et al. (2014)? It should not be understood only in terms of differences in brain activity. After all, one general conclusion supported by Schlagenhauf et al.'s (2014) study is that differences in brain activity observed between groups of mentally ill patients and healthy controls (and also within the same group of patients) are *not*, by themselves, reliable indicators of pathophysiological differences. This conclusion is independently underwritten by a clever study performed by Jonas and Kording (2017). They performed a number of interventions and data analyses on a micro-processor, aiming to understand the processor's stored-program computer architecture without any prior hypothesis about its (teleological) function. While Jonas and Kording (2017) concluded that the techniques they used are inadequate without a prior detailed analysis of tasks and the (neural and behavioural) responses they elicit, their conclusion is consistent with both mechanistic and semantic view of computation—more on this point in a moment.

Miscomputation should not be understood only in terms of differences in types of computational operations either. The types of computational operations examined in the study corresponded to a type of HMM, and to a type of Rescorla-Wagner model. Based on participants' neural and behavioural data, the former type of model was found to adequately fit the computational operations used by several schizophrenia patients, but also healthy controls; the latter type of model could capture the computational operations of the remaining schizophrenia patients. So, the kind of miscomputation uncovered in this study cannot be understood simply on the basis of differences in the *type* of computational operations involved in the task.

We should not understand the notion of *miscomputation* involved in Schlagenhauf et al.'s (2014) results only in terms of (mis)representation or lack of representational resources either. Consider representational resources first. Although a subgroup of schizophrenia patients might have used a computational strategy corresponding to a type of Rescorla-Wagner model that did not incorporate representations of task structure, this was evidence that pathophysiological differences cannot directly be read off from differences in computational strategy. It was not evidence that lacking certain representations could explain the observed abnormality in prediction error signalling between schizophrenia patients and healthy controls.

Consider misrepresentation now. Prediction error signals in RL models carry content about the difference between the difference between the learned predictive value of the current state and the sum of the current reward and the value of the next state. In particular, a reward prediction error signal δ(t) computed at time t is equal to r(t) + V(t+1) - V(t), where V(t) is the predicted value of some option at time t, and r(t) is the reward outcome obtained at time t. As Shea (2014) clarifies, the content of reward prediction error signals is both indicative and imperative, it has both accuracy and satisfaction conditions. The indicative content of the reward prediction error δ(t) is that the magnitude of the difference between the reward outcome r(t) and the predicted value of the outcome V(t) is proportional to δ. The imperative content of δ(t) is that the predicted value of the outcome V(t) should be increased (or decreased) in proportion to the magnitude of δ. Accordingly, δ(t) counts as a misrepresentation at least in two cases. First, when the magnitude of the difference between the reward outcome r(t) and the predicted value of that outcome V(t) is not proportional to δ(t). Second, when the expectation of reward V(t) is *not* updated in proportion to the magnitude of δ(t).

Now, a substantial body of work relying on results from computer science, neurophysiology and brain imaging indicates that phasic activity of dopamine midbrain neurons computes reward prediction error signals, namely: "their outputs appear to code for a deviation or error between the actual reward received and predictions of the time and magnitude of reward" (Schultz, Dayan, & Montague 1997, 1594; for a historical account see Colombo 2014a). This body of work connects a family of computational mechanisms for inference, integration, and choice to a neuromodulatory system implicated in several mental illnesses, the so-called dopamine reward system.

The accuracy and satisfaction conditions of reward prediction error signals refer to representational properties, specifically to the predicted value of a certain outcome. The content of these predictions, in turn, depends on individual dispositions like one's disposition to discount the value of reward outcomes as they approach a temporal horizon in the future, or to give more weight to losses than to gains with the same magnitude, which are captured by parameters in RL algorithms.

In Schlagenhauf et al.'s (2014) study, schizophrenia patients and healthy controls differed in model parameters capturing their differing sensitivity to observed rewards and their disposition to make a different choice on the next trial. Given these parameter differences—that is, given individual differences in expectation and outcome sensitivity—each participant's reward prediction error signals did not carry false content and played an appropriate role in updating individuals' representations of the expected value of different options. So, we should not understand the notion of *miscomputation* involved in Schlagenhauf et al.'s (2014) study as a kind of misrepresentation.

Let's now pay closer attention to the ways reliance on RL is justified in computational psychiatry, and, in particular, on what basis computational rules are ascribed to target neural mechanisms. The claim I want to defend is that *miscomputation*^{S}, but not *miscomputation*^{M}, adequately accounts for computational psychiatric practices similar to those involved in Schlagenhauf et al.'s (2014) study. The reason in support of this claim is that these practices always rely, at least implicitly, on something like a *specification* of the computing system of interest, and computational and representational functions essentially come together in any specification used by computational psychiatry. That is, any such specification type-identify computing mechanisms in the brain on the basis of the type of content carried by types of neural vehicles of interest.

Like in computer science and engineering (Turner 2011; Fresco, N., & Primiero 2013), specifications define what a computing system is expected to do at a certain level of abstraction. They define the problem a target system ought to solve, how it can solve it optimally, the set of admissible inputs and expected outputs, their content; they constrain the class of possible algorithms for solving the problem, and their adequate implementation. Specifications function as blueprints and reference documents for computer scientists, engineers, programmers and computer manufacturers and users. They allow for consistent communication about a certain type of system, and, most importantly here, defines when and to what extent a machine malfunctions. In Schweizer (2019, 41) words: "[i]t is only at a *non-intrinsic* prescriptive level of description that 'breakdowns' can occur, and we characterize these phenomena as malfunctions only because our extrinsic ascription has been violated."

In the case of brains understood as computing systems, different research communities in computational neuroscience and psychiatry employ different specifications, or, to employ Schweizer's (2019) phrase, they take different types of "computational stances" towards brain mechanisms of interest. These specifications are provisional, and can be revised on the basis of new evidence about the empirical adequacy of certain computational and representational ascriptions to a target mechanism, on the experimental accessibility of these ascriptions, on evidence about the physiological properties of that mechanism, on available technology for studying that mechanism,

and on theoretical considerations about the what it would take for that mechanism to make optimal use of its limited computational resources to solve a given task in a certain type of environment.

Any specification in computational psychiatry grounded in RL essentially appeals to semantic properties of the target system—for example, to the meta-representational properties of reward prediction errors putatively encoded by dopamine neurons—and does not typically appeal to functional, non-semantic, considerations grounded in the objective goals of survival and inclusive fitness or on the evolutionary history of mechanism of interest—for example, to the role of dopamine in human evolution. It is the semantic properties ascribed to a type of vehicle that contribute to determine the types of tasks that manipulations of vehicles of that type can solve, and what counts as "optimal" or "successful" performance in those tasks.

When different computational hypotheses are formulated about the same type of mechanism, this disagreement is empirical and can be resolved experimentally. After all, different content ascriptions to the same type of vehicle support different predictions concerning the rule-governed operations performed by the target mechanism in different tasks. In the light of empirical results, content ascriptions to a target type of vehicle can change, which can lead to novel computational specifications for a type of mechanism.

For example, as mentioned above, midbrain dopamine neurons are commonly thought to report a reward prediction error. While this content ascription has been highly successful in advancing our understanding of the mechanisms of learning and decision-making in both patients and healthy individuals, several lines of evidence indicate that dopamine activity encodes more dimensions of an error in prediction of an outcome unrelated to reward (Langdon et al. 2018). This body of evidence has been motivating novel content ascriptions—for example, the ascription that dopamine neurons report sensory temporal-difference errors, "encompassing both reward and non-reward features of a stimulus" (Gardner, Schoenbaum & Gershman 2018, 3).

These ascriptions really alter the hypothesised computational identity of a target system. If dopamine neurons signal errors in both sensory and reward predictions, then they would implement a type of RL algorithm that is different from model-free RL algorithms, which have commonly been thought to be implemented by the dorsolateral striatum and its dopaminergic afferents (e.g., Daw, Niv & Dayan 2005). As different semantic specifications of dopamine neurons alter the hypothesised computational identity of the dorsolateral striatum system, they motivate novel designs for testing competitive computational models. So, the specification-dependant, "perspectival" nature of content ascriptions does not need to involve indeterminacy or un-testability. It is rather a feature of historically situated research that promotes the pursuit of empirically adequate understanding of brains conceived of as computing systems.

Now, like most of the studies in computational psychiatry, Schlagenhauf et al. (2014) started by assuming brains are computing systems, and patterns of neural spikes are vehicles of computation. More specifically, they assumed the prefrontal-basal ganglia system carries out rule-governed operations over neural spikes, which support learning and decision-making. The reason in support of this assumption is that a body of convergent evidence from previous studies in a variety of experimental tasks indicate that the phasic responses of dopamine neurons in the basal ganglia represent temporal-difference reward prediction errors posited by RL algorithms. Schlagenhauf and collaborators did not ascribe certain rule-governed operations to their target system on the basis of purely functional considerations, since the selective history of the prefrontal-basal ganglia system and dopamine receptors are unknown or too uncertain (for reviews of the existing evidence see O'connell & Hofmann 2011; Yamamoto & Vernier 2011). Their computational ascriptions were based on the correspondence they justifiably assumed between certain types of computational states, as defined within a computational model, and features of their experimental task. The fact that tasks are semantically individuated does not logically entail that the computational system carrying out these tasks must also be individuated on the basis of its semantic properties. In practice, however, given existing knowledge of the evolutionary history of nervous systems, and of the dopamine system in particular, appealing to wide functional properties cannot not suffice to constrain, let alone to fix, computational ascriptions to prefrontal-basal ganglia mechanisms.

Given previous knowledge of what content dopamine vehicles might carry and given the reversal learning task they used in their study, Schlagenhauf et al. (2014) hypothesised determinate operations to the target system. Then, they presumed the neural and behavioural data they observed were generated by different types of operations, as defined by different algorithmic models of the task. A good fit between observed data and model predictions provided them with evidence for evaluating the empirical adequacy of their computational ascriptions, and for making normative judgements about the extent to which observed prediction errors are aberrant. This judgement relied on reasonable expectations embedded in a specification, which specifies how tokens of the type "reward prediction error" contribute to solve a given task reliably and efficiently.

The set of actual magnitudes of token prediction error signals in the human striatum is a relatively small subset of all possible prediction error magnitudes. In humans, the range of the mathematical function returning prediction errors is a relatively small subset of the function's co-domain. Appealing to the range (and also the sequence) of the magnitudes of token prediction errors in individuals from different populations, who try to solve a given problem, enable computational psychiatrists to characterise observed abnormalities in prediction error signalling as "statistically abnormal" or "suboptimal."

Claims of (sub)optimality depend on simulation results concerning the conditions under which a given RL algorithm converges to a global (or local) optimum solution for a given task, given a model of that task. These results set a normative standard against which human performance, either in healthy controls or in psychiatric patients, can be evaluated. They are also important to formulate hypotheses about possible computational constraints and parameterizations of algorithms that can explain human actual performance in a given task. Overall, they provide researchers with determine hypotheses about what a certain type of computing system ought to do in a given task.

In Schlagenhauf et al.'s (2014) study, the range of magnitudes of prediction error signalling in healthy controls was larger than in schizophrenia patients, who displayed reduced prediction error signals in the ventral striatum. Because of reduced prediction error signals, schizophrenia patients displayed blunted updates of the expected values of outcomes in a given state for future trials, which meant their learning was slower than healthy controls. So, in comparison to healthy controls, there was a greater divergence between what the prefrontal-basal ganglia system ought to have done in the given task and what it actually did. Their dopamine reward system did not perform as well as it could reasonably be expected.

What computational psychiatrists can justifiably expect for this type of neural computation in a given task depends on optimality results from computer simulations of the task, as I pointed out above, but also on existing knowledge of associations between a certain behavioural profile in a task and certain psychiatric symptoms. Although certain neural mechanisms of individuals with mental illness may carry out computations that solve a given task optimally, more efficiently or more reliably than computations observed in healthy individuals, their comparatively optimal computations would still be classified as symptomatic of *miscomputation* in some other task by many computational psychiatrists, since these optimal operations involve trade-offs in efficiency, reliability and timeliness of other computations in other tasks.[5]

To summarize, if the ascription of a *miscomputation* of prediction error signals observed in schizophrenia patients depends on a "specification", and this specification essentially involves the ascription of certain type of content to certain vehicles, and, on the basis of this ascription, determines what a target computing system ought to do without considering wider functional properties of the system, then at least one notion of *miscomputation* central to computational psychiatric practice requires a semantically laden characterization. Because the mechanistic account of miscomputation eschews semantics for individuating miscomputation, it cannot provide us with the resources to adequately explicate this notion of *miscomputation*. Whatever reason one may have to favour the mechanistic account over its competitors, that reason cannot be that the mechanistic

[5] Examples include the absence of optimism-biases in patients with depression, and the limitation of the positive self-reference bias in patients with social anxiety.

account does justice to the notion of *miscomputation* featuring in computational psychiatric practice relying on RL modelling.

5. Conclusion

I have argued that an explicatum of *miscomputation* grounded in the mechanistic view of computation, what I dubbed *miscomputation*^{M}, cannot do justice to relevant practices in computational psychiatry. I presented a case study representative of current approaches relying on RL and Bayesian modelling, and showed that these approaches appeal to a semantically laden notion of miscomputation grounded in a computational specification. If I'm right, what follows for the mechanistic view of physical computational?

One consequence is ontological and matters most directly for current debates about physical computation. The consequence is that there is no single, universally adequate account of the identity of physical computing systems (cf., Lee 2018). While everybody agrees that the operations of computing systems often consist of manipulation of meaningful vehicles (e.g., Miłkowski 2017), the disagreement concerns whether semantic properties play any role in the individuation of computation (see, e.g., Egan 2019). Now, one explicit aim of most accounts of the identity conditions of physical computing systems is to do justice to actual practices in the computational sciences. To the extent "doing justice to actual practice" is a descriptive, rather than a revisionary, project, these accounts should comport with successful scientific practices. If many successful practical and epistemic practices grounded in RL in computational psychiatry indicate that semantic properties play an essential role in enabling computational psychiatrists to determine the computational nature of a target system in the brain, then this is a reason to prefer some semantic account for the individuation of some types of computing systems. The mechanistic view, perhaps, better comports with successful practices grounded in in connectionist modelling, of which I said very little in this paper. In any case, if a mechanistic account of computational individuation does not comport with the successful practices I examined, then this fact counts against the account with respect to a certain type of system and practice.

Philosophers of computation who have tried to address questions about the individuation of physical computing systems have done much to raise the standards of clarity and explicitness in the statement of claims about the nature of concrete computing systems. But one substantive methodological question yet to be answered is: What role should information about scientific practice, including information about the concepts actually used by computational psychiatrists, play in an account of the identity conditions of physical computing systems?

One additional consequence of my argument concerns the adequacy of explications of the concept of (*mis*)*computation* grounded in the mechanistic view for illuminating certain types of

explanatory and methodological practices widespread in computational cognitive neuroscience. In particular, my argument provides additional evidence that concepts of (*mis*)*computation* that are not semantically-laden do not comport with explanatory practices involving requests for why it is that a target system *should* implement a certain kind of computational strategy to solve a given task (Chirimuuta 2014). While answers to these requests would ideally appeal to the teleological function of the system, as well as to its semantic properties, in practice teleological considerations play a limited role. What counts as successful performance in a given task, and why successful performance in that task depends on implementing a certain computational strategy are determined by a semantically-laden specification of the target system. This specification includes normative theories of learning and decision-making, whose justification does not lie in evolutionary considerations, but in theoretical results in philosophy, mathematics and computer science (Colombo 2019).

Acknowledgements

I am grateful to Dimitri Coehlo Mollo, Nir Fresco, Joe Dewhurst, and Corey J. Maley for 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

Barack, D. L., & Platt, M. L. (2016). Neurocomputational nosology: malfunctions of models and mechanisms. *Frontiers in psychology*, 7, 602. doi: 10.3389/fpsyg.2016.00602

Bechtel, W. (2012). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Routledge.

Brugger, S. & Broome, M. (2019). Computational psychiatry. In Routledge Handbook of the Computational Mind. Colombo, M. & Sprevak, M. (eds.). Routledge 468-484.

Cao, M., Wang, Z., & He, Y. (2015). Connectomics in psychiatric research: advances and applications. *Neuropsychiatric disease and treatment*, *11*, 2801-2810.

Carnap R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Chalmers, D. (2011). A Computational Foundation for the Study of Cognition. *Journal of Cognitive Science* 12, 323-57.

Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127-153.

Churchland, P. S. & Sejnowski, T. J. (1992). *The Computational Brain*, Cambridge, MA: MIT Press.

Coelho Mollo, D., 2018. Functional individuation, mechanistic implementation: the proper way of seeing the mechanistic view of concrete computation, *Synthese*, *195*: 3477–3497.

Cohen, J.D. & Servan-Schreiber, D. (1992). Context, cortex and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99, 45–77.

Colombo, M. (2019). Learning and reasoning. In M. Sprevak & M. Colombo (Eds.) *The Routledge Handbook of the Computational Mind* (pp. 381-396). New York: Routledge.

Colombo, M. (2014a). 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. (2014b). For a Few Neurons More: Tractability and Neurally Informed Economic Modelling. *The British Journal for the Philosophy of Science*, *66*(4), 713-736.

Daston, L. (1994). Enlightenment calculations. *Critical Inquiry, 21,* 182-202.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8, 1704–1711

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

Dewhurst, J. (2018) Computing mechanisms without proper functions. *Minds & Machines* 28(3), 569-588.

Dewhurst, J. (2014). Mechanistic miscomputation: a reply to Fresco and Primiero. *Philosophy & Technology*, 27(3), 495-498.

Egan, F. (2019). The Nature and Function of Content in Computational Models, in *The Routledge Handbook of the Computational Mind,* M. Sprevak and M. Colombo (eds.), Routledge, 247-258.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fresco, N. (2014). *Physical Computation and Cognitive Science*, Springer.

Fresco, N., & Primiero, G. (2013). Miscomputation. *Philosophy & Technology*, *26*(3), 253-272.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48-58.

Gardner, M. P., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B*, *285*(1891), 20181645.

Globus, G. G., & Arpaia, J. P. (1994). Psychiatry and the new dynamics. *Biological Psychiatry*, *35*(5), 352-364.

Huys, Q. J., Maia, T.V., & Frank, M.J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404-413.

Huys, Q. J., Guitart-Masip, M., Dolan, R.J., & Dayan, P. (2015). Decision-theoretic psychiatry. *Clinical Psychological Science*, *3*(3), 400-421.

Isaac, A. M. C. (2019). Computational thought from Descartes to Lovelace. In Routledge Handbook of the Computational Mind. Colombo, M. & Sprevak, M. (eds.). Routledge 25-38.

Jonas, E., and Kording, K.P. (2017). Could a neuroscientist understand a microprocessor? PLoS Comput. Biol. 13, e1005268.

Kemeny J. G. & Oppenheim P. (1952). Degree of factual support. *Philosophy of Science*, 19: 307–324.

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

Lee, J. (2018). Mechanisms, Wide Functions, and Content: Towards a Computational Pluralism. *The British Journal for the Philosophy of Science.* doi:10.1093/bjps/axy061

Maley, C. J. (2018). Continuous Neural Spikes and Information Theory. *Review of Philosophy and Psychology*, 1-21.

Miłkowski, M. (2017)- The False Dichotomy Between Causal Realization and Semantic Computation', *Hybris*, **38**, pp. 1-21.

Miłkowski, M., 2013, *Explaining the Computational Mind*, Cambridge, MA: MIT Press.

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

Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of science*, *58*(2), 168-184.

O'connell, L. A., & Hofmann, H. A. (2011). The vertebrate mesolimbic reward system and social behavior network: a comparative synthesis. *Journal of Comparative Neurology*, *519*(18), 3599-3639.

Patzelt, E. H., Hartley C. A., & Gershman, S. J. (2018). Computational Phenotyping: Using Models to Understand Individual Differences in Personality, Development, and Mental Illness. *Personality Neuroscience*. 1: e18: 1-10.

Pettersson-Yeo, W., Allen, P., Benetti, S., McGuire, P., & Mechelli, A. (2011). Dysconnectivity in schizophrenia: where are we now?. *Neuroscience & Biobehavioral Reviews*, *35*(5), 1110-1124.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

Piccinini, G. (2007). Computing mechanisms. *Philosophy of Science*, *74*(4), 501-526.

Rescorla, M. (2014a). A theory of computational implementation. *Synthese*, 191: 1277-1307.

Rescorla, M. (2014b). The causal relevance of content to computation. *Philosophy and Phenomenological Research*, *88*(1), 173-208.

Rolls, E.T., Cheng, W., Gilson, M., Qiu, J., Hu, Z., Ruan, H., ... & Zhang, X. (2018). Effective connectivity in depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 3:187–197.

Schwarzbold, M., Diaz, A., Martins, E.T., Rufino, A., Amante, L.N., Thais, M.E., Quevedo, J., Hohl, A., Linhares, M.N., & Walz, R. (2008). Psychiatric disorders and traumatic brain injury. *Neuropsychiatric disease and treatment*. 4, 797–816.

Schlagenhauf F, Huys QJ, Deserno L, Rapp MA, Beck A, Heinze HJ, Heinz A. (2014). Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *NeuroImage*. 89, 171–180.

Schultz, W., Dayan, P., & Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schweizer, P. (2019). Computation in Physical Systems: A Normative Mapping Account. In *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence* (pp. 27-47). Springer, Cham.

Shagrir, O. (2018). In defense of the semantic view of computation. *Synthese*. https://doi.org/10.1007/s11229-018-01921-z

Shea, N. (2014). Reward Prediction Error Signals are Meta-Representational. *Noûs*, *48*(2), 314-341.

Sprevak, M. (2010). Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science Part A*, *41*(3), 260-270.

Stephan, K.E., Bach, D. R., Fletcher, P.C., Flint, J., Frank, M. J., Friston, K. J., ... & Dayan, P. (2016a). Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. *The lancet Psychiatry*, *3*(1), 77-83.

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

Stephan, K. E., Baldeweg, T., & Friston, K. J. (2006). Synaptic plasticity and dysconnection in schizophrenia. Biological psychiatry, 59(10), 929-939.

Sutton, R.S. & Barto, A.G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tucker, C. (2018). How to Explain Miscomputation. *Philosophers'* Imprint 18 (24), 1-17.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433–460.

Turing, A. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceeding of the London Mathematical Society*, (series 2) 42(1), 230–265.

Turner, R. (2011). Specification. *Minds and Machines*, 21(2), 135–152.

Uckelman, S. L. (2018). Computation in Medieval Western Europe. In S. Hansson (Ed.) *Technology and Mathematics* (pp. 33-46). Springer, Cham.

Ursino, M., & La Cara, G. E. (2006). Travelling waves and EEG patterns during epileptic seizure: analysis with an integrate-and-fire neural network. *Journal of theoretical biology*, *242*(1), 171-187.

Wang, X. J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*(3), 638-654.

Yamamoto, K., & Vernier, P. (2011). The evolution of dopamine systems in chordates. *Frontiers in neuroanatomy*, *5*, 21.