Can an algorithm become delusional?

Broeker, Marianne D.; Broome, Matthew R.

DOI:
10.1007/s11097-023-09895-1

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Broeker, MD & Broome, MR 2023, 'Can an algorithm become delusional? Evaluating ontological commitments and methodology of computational psychiatry', *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-023-09895-1

Phenomenology and the Cognitive Sciences
https://doi.org/10.1007/s11097-023-09895-1

Can an algorithm become delusional? Evaluating ontological commitments and methodology of computational psychiatry

Marianne D. Broeker[1] · **Matthew R. Broome**[2]

Accepted: 4 February 2023

Abstract
The computational approach to psychiatric disorders, including delusions, promises explanation and treatment. Here, we argue that an information processing approach might be misleading to understand psychopathology and requires further refinement. We explore the claim of computational psychiatry being a bridge between phenomenology and physiology while focussing on the ontological commitments and corresponding methodology computational psychiatry is based on. Interconnecting ontological claims and methodological practices, the paper illustrates the structure of theory-building and testing in computational psychiatry.

First, we will explain the ontological commitments computational psychiatry is grounded in, the *Bayesian Brain hypothesis* (BBH) of unconscious inference, paired with normative deontic approaches applied to gauge psychopathology. We then turn to the steps taken in empirical paradigms, from definitions, which are used as starting points, to the operationalisation and isolation of cognitive processes and hypothesis testing based on algorithmic models, to consecutive interpretations regarding the aetiology of psychiatric disorders. We outline how experimental paradigms in computational psychiatry are specifically designed to confirm aberrations in assumed inferential processes, which are thought of as being the underlying *core invariant features*.

We will illustrate a gap between the ontological commitments of computational psychiatry and the operationalisation and testing of the cognition assumed to be relevant for psychopathology. This conceptual gap is of utmost importance when designing computational paradigms and may impede a crisp understanding of the approach. Lastly, in evaluating the conceptual gap, it becomes apparent that the information processing formalism used in computational psychiatry is still grounded in rational cognitive psychology.

Keywords Delusions · Predictive Processing · Bayes theorem · Deontic approach · Unconscious inference

Extended author information available on the last page of the article

Published online: 23 February 2023

1 Introduction

How similar are human mental processes to computational algorithms? In artificial intelligence (AI) research and in advances in cognitive simulations (CS), it has long been asked if and how human mental processes could be formalized and simulated with computational algorithms.

To formalize or simulate a mental capacity, we must know how inputs, i.e., incoming sensory information, can be transformed into outputs, hence, into the target cognition, and what algorithms could be used for that transformation. A specific algorithm captures more than the input-output mapping, since assumptions and constraints embedded into the input structure, meaning for instance characteristics of the incoming data, such as its distribution, and the agent are needed to make a hypothesis about an appropriate algorithm linking input to output, given the assigned constraints. Computational psychiatry operates in a similar way to these AI assumptions: It takes inputs from an experiment and behavioural outputs, summarized in parameters, and asks under what conditions as well as constraints the output could be created from the input. This is then formalized in an algorithm that becomes the hypothesis of the realisation of the input- to output mapping. Thus, mental capacities are defined by the mapping of input and output terms. Thereby, computational psychiatry aims to simulate unobservable mental processes, which are assumed to be algorithmic and underlie human cognition, thought, behaviour and psychopathology.

However, since the beginning of simulation research in AI, it has been questioned whether these transformation processes from input into output, when applied to human mental processes, are really of an algorithmic kind and can therefore be fully formalized. Dreyfus (1992), for instance, claimed that important human capacities, such as fringe consciousness or the broader insight structure of a problem cannot be formalized. Weizenbaum (1976) stated that love and respect are not technical problems. This builds on the idea that even if we would be technically able to formalize mental processes, that would not mean that this formalization would conceptually and meaningfully help us to overcome problems related to these capacities, as the solution might not be of a technical or mechanistic kind.

Computational psychiatry builds on the assumption that cognition and behaviour, can be explained computationally with algorithmic models, mapping input to output. While computational psychiatry is somewhat aiming to reform treatment approaches, more fundamental motivations target the formalisation of human cognition generally, and psychopathology specifically, which largely intersects with motivations of predicting and controlling human behaviour. As in Marr's levels of analysis (Marr, 1982), the computational level serves as a theoretical foundation, defining the mapping between an input, or initial state, and output, or resulting state, constituting a (i.e. mental) capacity. This highlights the dependency of the mental capacity or form of cognition on the incoming sensory information or other input structure. The algorithmic level describes the mechanism of how this capacity is realised and optimized, hence how the input is transferred into the output. Lastly, the implementation level states where in the physiological or technological system the capacity is realised. Whereas defining and mapping inputs and outputs is needed as theoretical embedding, the algorithmic level serves as the explanation of the mapping of input and out-

put that constitutes the phenomenon. Thus, computational models of cognition aim to uncover the mechanistic computations behind our emotions, thoughts, behaviours, and symptoms of psychopathology. Likewise, aberrations in algorithms are thought of as being the underlying causes and explanations of specific psychopathological phenomena.

Explaining cognition computationally means to describe phenomena mechanically in terms of inputs that lead to specific outputs and to find the algorithms of that transformation. In computational psychiatry, the realising algorithms are mainly, though not exclusively, thought of as being inferential. In terms of theory-building it remains to be seen to what extent computational psychiatry gives us new insights, rather than just being a formalized re- description of inferential processes.

The computational approach has been applied to various psychiatric and neurological disorders, i.e. positive, often isolated symptoms of schizophrenia (Adams et al., 2013; Corlett et al., 2019; Sterzer et al., 2018); autism (Lawson et al., 2014); Parkinson's disease (O'Callaghan et al., 2017); anorexia (Gadsby & Hohwy, 2019); addiction (Schwartenbeck et al., 2016), and depression (Barrett et al., 2016). As clinical delusions, now referred to as delusions, are one of the most commonly studied psychopathological phenomena in computational psychiatry (Adams et al., 2013; Corlett et al., 2019; Sterzer et al., 2018), they will be used as an example to illustrate how computational models are applied to psychopathology.

Inference generally is a method of (rational) reasoning based on evidence, whereas *Bayesian inference*, that is abductive inferences under uncertainty (Coltheart et al., 2010; Mathys et al., 2011), is mostly used in computational psychiatry and is a specific type of inferential computation. The role of uncertainty (Hohwy, 2013) as well as the notion of evidence to reduce that uncertainty are central to Bayes, where uncertainty of the outside world is dealt with through Bayes decision rule (Fahlman et al., 1983; Huys et al., 2011). Within the *Bayesian Brain hypothesis* (BBH), Bayes rule refers to the aim of the organism to predict what is happening next, in terms of a state in the world, a perception, an emotion, etc., where prior knowledge, hence internally generated predictions, are combined with new sensory evidence from the environment or the body, leading to new estimates (posterior) of what is about to happen. Thereby, the goal is to minimise the error resulting from the prediction, called *prediction error* (Clark, 2013; Friston, 2005; Hohwy, 2013). This has firstly been conceptualised with the notion of *free energy* as a measure of the discrepancy between actual features of the outside world and their internal representations (Hinton & Zemel, 1993), where action (active inference) and perception (perceptual inference) result from free energy minimization (Friston et al., 2006). If a prediction error occurs, it is typically transmitted back to an internal hierarchical knowledge system to update higher-level prior expectations, leading to dynamic interactions. Most theories on delusions focus on alterations in hierarchical inferences (Adams et al., 2013; Friston, 2005; Sterzer et al., 2018) across various interdependent levels of processing, starting from lower inferences at less abstract levels (i.e. perception of lower-level features) to higher-level inferences on abstract concepts (i.e. estimation of hidden world states). Two relevant hierarchical inference frameworks for delusion are *Belief Propagation/ Circular Inference* (Jardri & Denève, 2013) and *Predictive Coding*, including active

inferences (Adams et al., 2013; Corlett et al., 2019; Sterzer et al., 2018,). A detailed distinction about their characteristics can be found here (Ashinoff et al., 2021).

Predictive coding in conjunction with the BBH of delusions will further be used to illustrate *one* methodological approach in computational psychiatry, out of many other existing approaches. Conceptualizing the brain as a hierarchical prediction and information processing machine, the BBH aims to establish a mechanistic link between biological processes that implement information processing (Friston, 2010), quantitative computational models and symptoms of psychopathology. The assumption that cognition generally, and psychopathology, correspond to information processing as well as to aberrations thereof, represents an ontological assumption, or background belief (Maatman, 2021), about the nature of cognition and the causal structure of the world. It also lies within the deficit approach of psychopathology, where inferential aberrations and psychopathology generally are thought of as deficits. However, what constitutes an *aberration* in information processing must be defined. Thus, we need a framework that determines how information is *best* and *most optimally* processed, based on given inputs and outputs, to then establish aberrations thereof. Based on Marr's levels, theories on the computational level are often regarded as normative, e.g. rational or optimal normative (Oaksford & Chater, 2009). I.e. Bayesian modellers argue that computational theories are idealised optimization, serving rational goals (Anderson, 1990). Since Bayes rule is a rational, logico-mathematical rule, within the classical BBH, information processing follows a strategy where minimizing the prediction error is considered optimal. However, this Bayesian rule is only optimal if no other contingent factors are present. If, in a given situation or task, contingent facts are known to influence the relationship of input and output, what is optimal may become contingent on these other, known, influential factors. The optimal processing strategy, which is defined *a priori* in order to establish aberrations thereof, therefore also depends upon the structure and contingent facts of a given situation or task, which can be called *context-dependency of optimality* (Gigerenzer, 2008), or *embedded optimality*. Contingent factors could be the structure of the environment or task, limitations of the agent, or factors that have previously been considered "irrational" in comparison to processing in accordance with Bayes rule, such as processing costs of motivational influences (Williams, 2020). Importantly, when we define a framework of optimal information processing to map input to output, we assume to have knowledge of all contingent factors this mapping depends on, which then gives leverage to interpret what is most optimal.

Thus, optimality is defined *a priori*, in a fixed, general, and universally applicable way, as well as in reference to a causal structure of the world. In computational psychiatry, contingent factors are often controlled for, so that underlying Bayesian forms of rational processing can be laid bare.

Crucially, next to Bayesian and mathematical frameworks, other frameworks of rationality can be used, which then changes what is considered optimal or rational, for instance procedural rationality, epistemic rationality or agential rationality (Miyazono & Bortolotti, 2021). Lastly, an aberration in information processing can also be shown as *abnormal biases*, constituted by a statistical deviation (i.e., from a control group), without necessarily referring to an optimality framework.

Thus, the goal of computational psychiatry is to show aberrations in information processing algorithms, however, the aberrations depend on the way how optimal processing is predetermined which can change from study to study. After establishing aberrations from *optimal* algorithms and correlating them to symptomology, these aberrations are often assumed to underlie, hence, give rise to psychiatric symptoms (Griffin & Fletcher, 2017). Computational psychiatry assumes that at the computational level, symptoms can be re-constructed in an experimental task based on inputs and outputs, or simply simulated (Adams et al., 2015).

Importantly, the algorithms realising mental capacities are mostly conceptualised as operating indirectly, automatically and unconsciously (Mishara, 2007), as underlying mechanisms *behind* emotions, thoughts and behaviours. This dates back to early Bayesian inference theories, where perceptual inference was conceptualized as a basic unconscious mechanism (Helmholtz, 1925). Helmholtz (1925) further claimed that perception involves inferential processes somewhat conceptually similar to deliberate reasoning, an assumption often used to equate "*seeing is believing*", and to suspend the boundaries between perception, reasoning and beliefs. Thus, it has been argued that agents make rational but unconscious decisions.

Using the BBH as a theoretical framework implies several ontological commitments, meaning fundamental assumptions about the nature of cognition and psychopathology. Ontological commitments and auxiliary assumptions determine how we investigate and measure concepts, the proposed mechanisms, the confounds and the statistical analysis (Maatman, 2021). For one, there is the assumption that mental capacities consist of unconscious information processing algorithms and can therefore be formalized. This has also been called *psychological assumption* (Dreyfus, 1992). Next, it is assumed that psychopathology evolves around minimizing general uncertainty (free energy) of the immediate external environment, thus, pathology evolves in the realm of a shared single, external, objective world that entails uncertainties and hidden states of the world. Thus, uncertainty refers to a predefined situation or task rather than being a function of subjective needs. Thus, computational psychiatry entails a direct assumption that aberrations in free-energy minimisation are driving psychopathology. Further, it is assumed that psychopathology unfolds in rational decision-making processes, as these are mostly captured in the experimental paradigms and computational models. It is therefore assumed that rationality, in terms of an optimal strategy within decision-making, is a useful concept to gauge psychopathology. In terms of coherent theory-building in psychiatry, using the BBH with its ontological commitments connects a theory of psychopathology to a theory of the contingent facts that go into defining the optimal context and optimal information processing strategy, which are then used as an exhaustive explanation of the phenomenon. However, paradigmatically, we would need a theory of how minimising environmental structural uncertainty is related to psychopathology or delusions, for instance, how an intolerance of perceptual or decisional uncertainty is connected to symptomatology.

We will now discuss the methodological steps involved in the computational psychiatry of delusions.

2 Taxonomic considerations

*Classification has long been thought of to reveal the nature of a disorder as well as the structural features (*Kendell & Jablensky, 2003*).*

2.1 Taxonomy of origin (aetiology) vs. manifestation

Before a phenomenon can be studied empirically, or expressed in Marr's terms, and before the input and output mapping can be established, it has to be defined and classified. Both phenomenology and computational psychiatry typically search for *core invariant features*, meaning underlying characteristics of a phenomenon, such as particular symptoms, which then typically become the criteria by which a disorder is classified. Since the notion of core invariant features assumes that a disorder has defining characteristics, it unfolds within realism, though the characteristics can, but do not have to be of a natural kind (Broome, 2007). Within realism, it is assumed that the classifications are translatable into testable empirical sciences (Mishara & Sterzer, 2015; Parnas & Zahavi, 2002).

In contrast to that, anti-essentialists assume that classifications fulfil certain purposes, such as reflecting value-judgements (Thornton, 2002) but disorders do not necessarily have a discrete essence (Agich, 2002; Horwitz, 2002; Skene, 2002). Thus, since there is no essence, classification is driven by theories and values (Cooper, 2004; Haslam, 2002). Phenomena are seen as perspectival, incomplete, and infected with interests (Sadler, 2004). Classification itself, within anti-essentialism, is therefore distinct from nosology and most importantly from the assumed aetiology of a disorder. Thus, from an anti-essentialist point of view, using taxonomy as a starting point for empirical inquiry is misleading, as it does not represent core invariant features.

Computational psychiatry research typically starts with a definition of a phenomenon that it wishes to explain. In recent years, definitions used as starting points for research have moved towards DSM-5 based classification (behavioural criteria), in which delusions are defined as *thoughts held with rigidity and certainty in light of contrary evidence* (Apa, 2013)). The focus has been shifted from earlier definitions including broader self-concepts (such as sense of agency, minimal self) to delusions as a thought-based phenomenon, which is called doxastic account.

DSM categories, specifically the symptoms listed under the DSM categories, are often used as starting points to define the phenomena for which a computational explanation, hence model, is being sought out. Thus, the unit of investigation often is the (DSM) disorder or specifically listed DSM symptoms, such as rigid beliefs. It is important to emphasise what DSM taxonomy, hence DSM diagnosis and the listed symptoms, represent and how these come into existence. Crucially, the formation process of DSM diagnoses underlies a descriptive operationalism approach, which is not always scientifically driven, but results from conventions and intersects with hidden motives, political, economic, social, and pharmaceutical interests, and values (Ghaemi, 2009; *Pickersgill,* 2014*). It is this interaction with various other interest groups and the operationalism approach which makes it questionable to what extent DSM diagnosis and symptoms represent valid scientifically proven (disease, bio-*

logical, social, etc.) entities and can thus be the starting point for further academic endeavours. For instance, the American Psychiatry Association (APA), which produces the DSM, receives substantial pharmaceutical industry funding, panel members of the DSM committees hold drug industry ties (Cosgrove et al. 2011, 228–32). Generally, the influence of industry, especially pharmaceutics, has been said to have led to an expansion of psychiatric diagnosis and over-diagnosing 'nosologomania' (Ghaemi, 2009), as well as the 'medicalisation of normality' (Pickersgill, 2014). On the scientific level, the interference of different interest groups can lead to ill definitions of clinical and research subject groups. In sum, the DSM and science generally are always embedded into a greater socio-economic context, which heavily influence the dominant paradigms and makes it extremely unlikely that DSM diagnoses correspond to specific disease-entities, for which underlying computational mechanisms can be found.

Given the realism approach of empirical science, taking rigidity and certainty of thoughts as defining characteristics (e.g. Baker et al., 2019), gives them the status of core invariant features of delusions, though not all paradigms define delusions in terms of rigidity and certainty (e.g. Bansal et al., 2022).

Alongside the emphasis on rigidity and certainty of thoughts comes the focus on the form, the structural criteria, rather than the content or meaning of delusions, which have also been proven important for delusions (Ritunnano et al., 2022). The form of phenomena, thus, the structural criteria refer to formal aspects, as *how* something is or can be described from the outside, rather than *what* it is, what content it entails. In other words, structural, hence formal features refer to a category of phenomena generally, without specifying the specificities that each lived experience has individually.

In this sense, the form becomes the core invariant feature. Furthermore, by referring to the form, rationality and optimality directly relate to the structure. Akin to an algorithm not understanding the meaning, computational psychiatry is mostly concerned with the structure. Focusing on the form rather than on content and meaning not only makes the definition content- and context-neutral, but it also makes the classification quantitative and structural. As criticized by Williams and Montagnese (2020), abnormalities are suggested to be domain-general, ranging over all possible contents of thought and all delusional types and themes.

Given the focus on the form as a definition and starting point for empirical explanations of delusions, it can now be asked if the form of a phenomenon represents the origin or rather a manifestation of it. Thus, if the form of a phenomenon can be a meaningful explanation of that very phenomenon. Importantly, in contrast to the DSM being only descriptive and not asserting aetiological or ontological claims, computational research takes DSM definitions as starting point not only for formalized re-descriptions, but also for the explanation of a given phenomenon, i.e., delusions.

To summarise, while research definitions of delusions have become increasingly based upon the DSM, computational research on delusions has endeavoured to focus on the form rather than the content. However, there are newer paradigms that do focus on the content, which will be addressed later on. While taxonomy does not look for an origin of a disorder and does not make ontological claims but can be arbitrary or

practical in its classifications, the goal of empirical research, including computational research, is often to find the apparent *cause*, hence explanation of a phenomenon.

Using taxonomy, in particular the formal structures of a phenomenon as the starting point of aetiological endeavours illustrates that computational psychiatry runs the risk of confusing a manifestation for an explanation or origin. Thus, it should be asked if, and under which premises, the form of a phenomenon can lead to an explanation of the origin and cause. This question is tantamount to asking what the shape of a raindrop can tell us about the causes of rain.

Importantly, formalization in general requires adequate and concise definitions of the constructs,entering the models as input and output, as robust phenomena (Eronen & Bringmann, 2021). Here, it could be questioned to which extent DSM classifications represent robust and concise phenomena, rather than conventional categories with loose boundaries (i.e., spectra).

3 Rationality as *modus operandi* of the mind

3.1 Beliefs as central parts of delusion categorisation and their appeal to (ir) rationality

The mind has long been conceptualised as *rational*, in philosophy as well as in psychology and cognitive science (Ellis, 1957). In a broader philosophical sense, rationality refers to the coherence of human thought in a structurally and intersubjectively understandable way. However, in computational psychiatry, rationality refers to a predefined optimal. This forces the question to what extent understandability and optimality are conceptually related.

Central to the definition of delusions are *beliefs*, which have been ascribed different attributions, from being *false* in DSM 4, 5 (Apa, 1987) to being *irrational*, as well as *rigid* and *certain*. These attributes typically gain their meaning in reference to shared evidence and shared frameworks of meaning, that stipulate how to relate to that evidence, which can be called *propositional framework* (Wittgenstein, 1974), or shared and single objective reality, which possesses a causal structure. Thus, delusional beliefs are typically characterised in reference to shared, external evidence. It should be noted that there are non-doxastic accounts of delusions that explicitly do not refer to delusions as beliefs, but i.e. as attitudes towards mental acts (Stephens & Graham, 2004). It has been outlined by Bortolotti and Broome (2008) that (ir)rationality as such is not sufficient to explain delusions, as (ir)rationality has been shown in all humans, i.e. in deductive, statistical reasoning (Samuels et al., 2002; Stanovich, 1999) as well as in syllogistic reasoning. Furthermore, various cognitive biases to deductive reasoning (i.e. confirmation bias), have been shown in healthy populations as well, so that their causal role for delusions can be ruled out (Sullivan-Bissett & Noordhof, 2020). However, the aforementioned authors referred to (ir)rationality mainly as logic or deductions. Whereas, as we will see further below in the chapter, (ir)rationality can refer to logical (ir)rationality as well as to probabilistic (ir)rationality, which are two different types of mathematically defined (ir)rationalities that can be applied to delusional beliefs. In recent years, the conception of the mind as a logi-

cal rational engine (Piaget, 1957), has been replaced by the conception of the mind as a probabilistic rational engine (Oaksford & Chater, 2009).

Concepts such as rigidity and certainty, that do not seem to explicitly refer to (ir) rationality at first, can appeal to the notion of rationality in a probabilistic sense. One form of probabilistic rationality is Bayesian rationality. The difference between logical rationality and probabilistic Bayesian rationality, as stated out by Oaksford & Chater, (2009) is that the former involves a monotonicity assumption, meaning that the inferential relations are held with certainty, are truth-preserving, and contingent facts cannot be accommodated. Whereas within the probabilistic Bayesian rationality approach, rationality is defined as the ability to reason about uncertainty, while accommodating contingent facts. This non-monotonicity seems to suit the everyday world better than an absolute certainty assumption. Within non-monotonicity, any conclusion can be overturned if more information is acquired, as inference itself is uncertain. Furthermore, probability itself refers to a degree of belief rather than to objective facts. In short, probability theory seems, as stated by Oaksford & Chater, (2009), much more suitable to deal with the non-monotonic, uncertain character of everyday reasoning.

However, although probability theory does not assume absolute certainty, it assumes certainty to a specific degree of precision. Thus, it assumes *truths* that are more or less likely and thereby builds on the same operational premises as logic, while taking away the absolute certainty of the assumptions made. Within non-monotonicity, statistical or probabilistic truth gets the pretence of absolute truth. This also becomes apparent in the juxtaposition of deductive to abductive and inductive inference, whereas the former refers to premises as being true and conclusions being certain, the latter one refers to premises being possibly true and conclusions being probable. In computational psychiatry, probable truth is treated as absolute *truth*, although it only represents precision. In sum, within a mathematical approach to beliefs, the rationality of beliefs can either be logical or probabilistic, however, in the Bayesian Brain hypothesis, rationality is thought of as being the latter.

3.2 The harmony of the computational model and empirical paradigm

After defining the phenomenon that is to be modelled (i.e., *rigid, and certain beliefs*), it needs to be outlined how a mathematical theory (i.e., logic, probability/Bayes) can be connected to the empirical data from the behavioural- or model-based measures obtained in an experimental task. The model-based measures are typically captured in the parameter estimates of the computational model to quantify the behaviour. Further, the model-based measures are usually designed to align with the definition of the examined phenomenon, i.e., *rigid*, and *certain* beliefs. Thus, the definition directly influences operationalisation and measurement. Parameter estimates, hence, behavioural measures are collected in tasks that, in the case of delusions, usually though not always, isolate probabilistic reasoning problems, e.g., reasoning about a hidden state or estimating a probability. Thus, computational paradigms on delusions mostly operationalize and isolate active reasoning behaviours in tasks like the beads task or other tasks targeting probabilistic thinking. In that way, the isolated behaviour

(i.e., a type of reasoning) can be selectively attributed and linked to altered inferential processes, as well as to specific symptoms.

After collecting the model-based measures, a computational model links the input, information given and relevant in a task, and output, parameter estimates representing the behaviour. This algorithmic level provides the hypothesis for the mapping and is the level where aberrations relevant for psychopathology are assumed to reside. Importantly, the type of cognition assessed, or behaviour isolated in these tasks is often of an explicit, deliberate, and conscious kind. Using the distinction made by Shea and Frith, (2016) into type 0, 1 and 2 cognition, Type 0 cognition represents automatic processing and unconscious representations, whereas type 1 represents automatic processing and conscious representation, and type 2 represents deliberate processing and conscious representation. With this classification, the isolated behaviour in most, though not all, computational task design on delusions (Corlett et al., 2019; Reed et al., 2020; Sterzer et al., 2018) represents type 2 cognition, which is explicit, deliberate and conscious, and thus fundamentally different from unconscious and automatic types of cognition.

Further, the algorithmic level simulating the underlying processes is mainly based on isolated decision-making, hence, type-II cognition. Thus, the design of inferential reasoning tasks seems contrary to the theoretical assumption that Bayesian inferential processes take place automatically and unconsciously. If Bayesian inferential processes were to take place unconsciously, most current computational paradigms, isolating deliberate reasoning and decision making don't capture unconscious and automatic processes.

For reasoning tasks (e.g. the beads task, i.e. to elicit jumping to conclusions or data-gathering biases, that is often used in computational psychiatry), Oaksford & Chater, (2009) suggested that normative analysis, as a component of rational analysis, can be used to connect mathematical models to the observed behaviour, leading to explanatory accounts. Rational analysis (Anderson, 1990) is a way to capture and describe empirical data concerning thoughts and behaviour, as well as the optimal way for these behaviours to unfold. When using a normative approach, computational psychiatry applied to delusions (mainly) uses Bayesian rational analysis to describe the empirical data. Rational analysis is a procedure that follows several different steps: (1) Specification of a precise goal of what the cognitive system wants to achieve; (2) Specification of a formal model of the environment with specific constraints (including the structure of statistical regularities); (3) Specification of the constraints of the organism. (4) Consecutively, optimal behaviour is derived from steps 1–3, which requires using formal rational norms, such as probability theory (Anderson, 1990). Using Bayesian analysis as a modelling frame defines the optimal way to map input to output, depending on predefined goals and contingent facts. Importantly, what makes rational analysis *normative* is the definition of optimal behaviour, hence, an optimal algorithm, depending on predetermined goals, constraints, as well as contingent facts.

Bayesian rational analysis appeals to behavioural tasks as deontic tasks. These are tasks that express rules of conduct and norms rather than facts about the world, as in logic. In deontic selection tasks, conditionals describe rules not how people neces-

sarily behave, *but how they should be behaving (inferentially optimal)*, regarding a specific, predetermined goal.

The rule is not a hypothesis under test, but a regulation that *should be obeyed* (Manktelow & Over, 1991). The conditionals now do not concern veracity and therefore can neither be confirmed nor disconfirmed by any observations of actual behaviour (Oaksford & Chater, 2009). The *law* is that people ought to behave to maximise expected utility, hence achieve an a priori stipulated goal or minimize uncertainty. Rational analysis then aims to detect violators from the law, rather than finding the truth. What further underlies the deontic approach is the assumption that the brain operates in evolutionary nearly (Friston, 2010) or approximately (Williams, 2020) optimal ways, which is expressed in the *ideal observer*, or in philosophy, this has been called idealized rationality.

Crucially, the deontic and normative approach represent what is claimed to make Bayesian rational analysis more than a mathematical re-description of the behaviour, as it describes *why* a particular algorithm is used to solve a predefined problem and how it might be optimized. The problem or goal is determined by what we know about the environment and the agent, in other words, the input or sensory evidence structure and noise. A goal or utility is then extrapolated from that structure to determine how the agent ought to behave. A specifically fixed and predetermined goal or value attached to specific contingencies is also called *utility function* in computational modelling. Thus, the optimal algorithm arises from the problem itself, the problem drives the solution (Anderson, 1990). This could also be called: *problem-dependency or goal dependency.*

Within theory-building, in regard to the problem-dependency of the algorithm, representing the explanatory level, we need to identify whether the problems and input structure stated in tasks where rational analysis is applied to (i.e., beads task, conditional reasoning task), structurally mimics real-world problems as well as problems relevant for the symptoms of psychopathology. In the following, two accounts will be described and critically evaluated that argue for the explanatory value of computational psychiatry:

3.3 Failure modes as deviations from optimal algorithms – 1st explanatory, aetiological account of computational psychiatry

Computational psychiatry, building on rational analysis and the deontic approach, is explicitly normative, which means that it describes psychopathology in terms of aberrations from what is considered *optimal* cognition or behaviour. Thus, rationality, which is now optimality, is used as a benchmark to gauge psychopathology. Normative models move from pure description to an apparent explanation of *why* a phenomenon occurs by looking at optimal behavior, under specific constraints and in relation to specific goals. While optimality refers to the optimal algorithm underlying that behavior, deviations or aberrations from the optimal algorithm are considered to underly psychopathology. These deviations are called *failure mode* (Ashinoff et al., 2021; Redish et al., 2008), mostly relating to form, not the content, as stipulated in the definition and subsequent operationalisation. Crucially, failure mode means what deviates from a stipulated norm, regardless of whether it is less or even more

optimal (Baker et al., 2019). Further, underlying failure modes are thought of as leading to specific behaviours and phenomenological states related to psychopathology, i.e., delusions. In addition, failure modes give a computational explanation for the observed and as aberrant classified task behaviours. For instance, specific behaviours, such as *increased information sampling*, can, through the normative, deontic approach be identified as underlying inferential failure mode (i.e., *here overweighting of priors at different hierarchical levels*). Hence, to label something as failure mode, the normative model needs to act as a benchmark to gauge psychopathology. Further, the failure mode is accompanied by an explanation of what led to that failure mode (e.g., *under- or overestimation of noise or environmental volatility*), which then becomes the central explanation of a (pathological) phenomenon. Depending on the input structure and specifically set goals, which can enter models as utility functions, different failure modes can be determined. Failure modes are therefore completely contingent on predetermined fixed goals, and do not allow for flexibility.

Importantly, also a utility function adds a fixed goal, rather than flexible one, as it adds further determined facts to the already determined facts, which are then often called *values*. In other words, per-determined values are embedded into the experiment at different stages.

There are simple *Bayesian optimal models* (*ideal observer*), which do not specify specific constraints, and there are more specific models. Given a specific normative model, what is defined as optimal and contrarily as failure mode will vary based on the internal or external constraints as well as the consecutively stipulated goals: In prescriptive or functional models, parameters vary as a function of given external or theorised internal constraints (e.g., noise) in information processing. The stipulation of goals under constraints gives prescriptive models the ability to deliver not only a definition of optimal but also, contingent on the goals, a mechanistic explanation of the *cause*. There are many prescriptive models, e.g., the volatility uncertainty and bounded rationality account.

3.3.1 Volatility environmental uncertainty account

The structure, including uncertainty embedded in the environment, which is recreated in the experiment as input structure, determines the goal and consecutively what is considered optimal. In this account, the important dimension is environmental volatility, which is experimentally operationalised as the frequency of unannounced changes, i.e., in hidden states. Depending on the specific model constraints, over- or underestimating environmental volatility becomes the optimal behaviour or the failure mode and thereby the explanation for a given deviant behaviour. Empirically, within computational accounts of delusions, this model has been used in social and non-social environments (Reed et al., 2020). However, the explanation by Reed et al. (2020) has been criticised as it is less clear how greater expectations of environmental (social and non-social) volatility are connected to the specific content of paranoia or persecutory delusions (Williams & Montagnese, 2020). However, Reed's hypothesis explains the form of delusions (rigidity and certainty), which explicitly excludes the content.

3.3.2 Internal noise (bounded rationality model)

In this model, optimality depends on the noisy neural sample, meaning an internal representation of the prior belief. Optimal prior weighting is governed by the internal cost of improving the precision of these representations. Thus, prior weighting is an adaptive response that depends on the internal constraints of information processing. Prior overweighting as *failure mode* might result from alterations in prior sampling or the strategies used to resolve the trade-offs.

Looking at empirical findings, it seems less clear whether persons with delusions exhibit any significant domain-general inferential impairments that could be explained with an algorithmic failure mode, as studies trying to identity failure modes have been inconclusive (Ashinoff et al., 2021), and typically identify diverse rather than one coherent failure mode. Different failure modes, connected to aberrant (inferential) reasoning behaviours, have been proposed and consecutively discounted, such as the JTC bias (precision-overweighting of sensory information) (Adams et al., 2013; Corlett et al., 2019; Sterzer et al., 2018), which seem to be related to schizophrenia, but not to delusions in particular. Although, also the relation to schizophrenia has been explained through non-inferential processes, such as general cognitive impairments (Ashinoff et al., 2021; Tripoli et al., 2020) which have often not been controlled for in meta-analyses (McLean et al., 2017).

3.4 Measuring unobservable states – 2nd explanatory, aetiological account of computational psychiatry

Input and output structures are directly observable; the input is given and controlled for in the experimental task and the output is measured in the response behaviour. The algorithmic level, which simulates the cognitive process transferring the input into the output, is, however, not directly observable but can be inferred from the input and the behaviour, with various parameters. This illustrates that the algorithmic level is still dependent on the input and output structure, it is a hypothesis of how the input and the output are optimally connected. Computational psychiatry assumes its algorithmic models represent unobservable states, hence underlying implicit processes that the brain performs to produce the behaviour, but which are still distinct from the behaviour. This algorithmic hence inferential level is mostly assumed to be unconscious and automatic.

At this point, some concerns about the ontological commitments of computational psychiatry can be raised: Firstly, it can be asked if the underlying cognitive process that transfers input into output really simulate something that takes place automatically and unconsciously, or if the cognitive process isolated in the task is rather a formal re-description of explicit and deliberate reasoning.

An underlying psychological process, even if explicit and deliberate, is still not directly observable and still distinct from the explicit output, the behavioural result. Therefore, it follows that only because the process itself cannot directly be observed and measured, it does not make the (underlying) process any more or less explicit, conscious, or deliberate. Thus, whether the transformation of the input to the output is a type 0, 1, or 2 cognitive process is determined by the task itself and the cogni-

tion isolated in it. Even if there might be other processes contributing to a conscious and deliberate reasoning task, which might be automatic and below awareness, the targeted cognition is still the reasoning or decision-making process itself. Thus, depending on the isolated behaviour in a task, the computational model is simulating the underlying process of that very isolated behaviour or cognition. Thus, it should be stipulated based on the isolated task behaviour or cognition itself, whether the underlying process represents type 0, 1 or 2 cognition, perception, or a reasoning-based process. The underlying process might not necessarily be unobservable for the individual, but it is unobservable for the experimental observer. To gauge whether the algorithm simulates something automatic and unconscious, it should be asked whether the type of cognition the algorithm is simulating is more explicit or implicit, and if that type of cognition can be captured based on input and output. For explicit behaviour for instance, an algorithm might capture processes that are themselves explicit but not directly observable. The algorithm goes beyond the observable, but this does not make the simulated process itself any more or less conscious.

Lastly, even if conceptualised as unconscious and implicit, the assumed underlying algorithmic mechanisms represent an explicit process in itself. Computational processes in themselves, due to their stepwise and rule-based nature, inherently represent explicit processes. Thus, it could be asked whether the idea of implicit underlying processes should perhaps be detached from explicit computations. Furthermore, there is also very little reason to believe that unconscious processes follow explicit, step-by-step rules, as these mostly require deliberation. Now, it seems less surprising that algorithms are particularly good in simulating higher-order functions, such as reasoning.

In terms of theory- building, if we were to appreciate the role of higher-order reasoning in psychopathology, we would need a characterisation and aligning background theory of how reasoning and higher-order cognition (in decision-making) relates to pathology.

To summarize, predictive coding models assume that human cognition is driven by reducing uncertainty, which can be uncertainty of internal representations or of external, environmental structures. Thereby, normative models stipulate the optimal algorithm to establish a *mechanistic* relationship between input and output. Importantly, uncertainty is assumed to influence every individual in the same way, it has a fixed, objective, "ontic" and "deontic" structure. As uncertainty resides within a fixed structure, and relates to a causal structure of a single, objective reality, it does not arise for individuals in an independent way and it is deprived of content, context, meaning as well as flexibility. What is rational, and in the case of computational psychiatry optimal, and what is a failure mode, therefore directly depends on the structure of the internal or external uncertainty and the goals contingent on that structure. Rationality, hence, optimality, refer to managing uncertainty in a given evidence structure, rather than a meaning. Understandability becomes of a mathematical kind.

In sum, a model can be seen as a hypothesis and simulation of the aetiological process of a disorder or phenomenon, i.e., delusions, based on the failure modes on the algorithmic level, which are assumed to be implemented within a biological structure. Lastly, specific deviation from optimal, hence failure mode, are linked back to specific clinical phenomena by correlational analysis (Baker et al., 2019), typically

with clinical questionnaires, which also mainly focus on structural criteria, in fewer cases, causal analysis (Suthaharan et al., 2021), using clinical interventions, such as medication or TMS, to measure change in parameter fits.

For rational analysis generally two caveats have been pointed out by Oaksford & Chater (2009), which will be related to computational approaches of psychopathology: (1) Bayesian rational analysis generally is not intended to be a theory *of* psychological processes, which questions its suitability as a theory of psychopathological processes. This means that it does not specify the processes or algorithms that are actually used to carry out the solution, as those can take many different forms (Anderson, 1990). Instead, processes arise because the cognitive system is well adapted to solving a particular problem. In short: Bayesian rational analysis explains the rational reasoning processes that are isolated in experimental designs, but not the operating processes that underly rational reasoning.

(2) Understanding the structure of reasoning (i.e. from a Bayesian or logical perspective), should be distinguished from measuring people's performance on logical or inference problems (Kahneman et al., 1982). Behavioural performance on logical or probabilistic problems results from explicit application of instructions and reasoning, rather than illustrating the capacities and structures immanent to the mind. Even if the mind would be a computational organ, it's not possible to engage this machinery with verbally or numerically stated probabilistic tasks (Oaksford & Chater, 2009).

Thus, it becomes questionable to what extent probabilistic reasoning tasks should be used to evaluate underlying implicit and unconscious mechanisms, especially since the evidence collected in most computational paradigms illustrates patterns of qualitative type-2 reasoning that people find natural. Bayesian rational analysis, as mostly applied to computational paradigms on delusions, provides a rational analysis of human reasoning (as isolated behaviour) however, in alignment to the argument by Oaksford & Chater (2009), it remains questionable how claims about underlying computational mechanisms could be derived from these kind of experiments.

Most tasks in computational psychiatry claim to test underlying implicit inferential mechanisms, which would be captured in type-0 cognition, automatic and unconscious processes. However, what is collected in the tasks, such as the beads task, are qualitative, deliberate, explicit patterns of reasoning, hence, type-2 cognition. This becomes evident in biases unfolding around processes such as evidence gathering and integration of evidence or belief evaluation and shows the inherent connection of computational psychiatry to cognitive psychology. Thus, the computational Bayesian approach is still based on rational analysis and therefore technically places psychopathology within the realm of a reasoning problem or disorder of thought (disorder of beliefs). This seems quite contradictory to the *one-factor theory* (Corlett et al., 2019; Sterzer et al., 2018), equating perception to higher-order processes, by lifting the distinction between them and assuming that the same mechanisms underlie both, based on the Helmholtzian claim. This then leads to the claim that either perception or reasoning can experimentally be looked at, as both result from the same underlying cognitive mechanism. However, on the functional and experiential level alone, it could be questions to which extent perception represents higher-order functioning and vice versa. In other words, equating perception with higher-order cognition only makes

sense on the computational level, which remains hypothetical, and which applies the same algorithms to phenomena ranging from perception to planet trajectories.

In conjunction with the *one-factor theory*, it should be noted that there are more recent attempts by Corlett and colleagues (Bansal et al., 2022; Rossi-Goldthorpe et al., 2021) to address the limitations of the beads task or similar experimental paradigms that mainly target and isolate deliberate reasoning related cognition, and instead to focus on perceptual cognition, actively bypassing deliberate reasoning, as the accounts assume that pathology manifests in implicit processes. Furthermore, there are experimental studies using a Bayesian framework that deliberately includes *irrational*, hence non- Bayesian, and content related factors, such as motivational biases or social vs. non-social influences, contributing to a cognitive outcome. For instance, the study by Rossi-Goldthorpe et al. (2021) used a perceptual classification task, where not the classification as scene vs. image, which still represents a deliberate and explicit albeit perceptual decision, but the switches in decision under social influence was the primary variable of interest, in conjunction with confidence in one's own and a social collaborators or competitors' choice. Paranoia was thus operationalised as a function of higher self-deception (distrusting one's own but trusting someone else's judgement) and over-confidence, in the context of a perceptual classification task. Self-deception was, in the context of this study, operationalised as something that is *irrational*, but can nevertheless be explained in Bayesian terms. To evaluate whether the claims of the study hold depend to a significant extent on the operationalisation of paranoia as self-deception, its construct validity as well as its ecological validity in reference to clinical phenomena, and their connection to reducing externally given uncertainty. Further, this example shows how social vs. non-social influences, as content, are integrated into the paradigm, where modelling is used to quantify the social influence vs. the influence of external structural uncertainty on the Bayesian decision process. Thus, it illustrates the integration of circumscribed, predetermined content-related factors that enter separate models based on group affiliation. Lastly, self-deception represents a specific form of irrationality, where generalisations outside the experimental context should be made carefully. Although the variables of interest were only dependent on but did not directly represent cognitive mechanisms, the underlying ontological commitment still seems to be that paranoia arises within decision-making processes, which again represent active and explicit cognitive mechanism.

Another example where deliberate conscious reasoning components have deliberately been bypassed is the study by Bansal et al. (2022), which used a belief-updating paradigm in a simple perceptual decision task to dissociate belief-updating processes in delusions at perceptual and higher cognitive levels. While these studies go in the right direction by moving away from isolating deliberate reasoning, and the random-dot motion paradigm used in this study already represented a less complex cognitive capacity, the judgement of direction in a dot-paradigm still represents an active evaluation and decision-making process, a form of cognition that is, at closer inspection, less of a perceptual, lower-level kind.

In summary, paradigms utilising algorithmic models, since the *optimal* solution is problem-, dependent but context- and content-independent, it should be asked whether the problems given in these experimental tasks resemble problems that

play important roles in every-day life as well as in psychopathology. The normative approach claims to make computational psychiatry more than a mathematical re-description of a psychopathological symptom, e.g., delusions. However, for it to be more than a re-description, it needs to assume universal behavioural *laws*, which are contingent on the chosen model. These deontic laws thereby replace facts but have a similar status after claiming to establish not how people actually act, but how they ought to. Having a law-like, normative status, they appeal to the truth in a similar way, although a probability is now *enough* to establish truth, rather than proving something with absolute certainty. Furthermore, these laws abide to a mathematical or otherwise fixed and predetermined optimal solution, which is entirely contingent on the constraints, hence the statistical regularities of a given input structure. Going back to delusions, it should be asked whether a deontic law is a useful concept for psychopathology, if psychological health can be described as obeying some ideal norms of rationality, describable in mathematical ways. Or in other words, whether *(ir)-irrationality* ascribed to psychopathology is of an inferential and formal kind.

3.5 (Synthesis) points of paradox

Based on the discussion above, we will now summarize the key points of tension in the computational approach and how they could be reconciled.

Whereas the idea of the brain as working in inferential way was initially meant to describe automatic and implicit perceptual processes (Helmholtz, 1867, 1925), which has been extended to the claim that delusions and hallucinations are similarly based on automatic and unconscious inferences (Adams et al., 2021; Mishara, 2007), most computational paradigms on clinical delusions operationalize and isolate deliberate reasoning, which relies on type 2 cognition, is explicit, deliberate and routed in rational theory and cognitivism. Computational psychiatry thus seems to represent a fusion of ideas, conceptually merging cognitivism, and rationalism with automatic, unconscious inferential Bayesian accounts. However, conclusions for type 0 unconscious cognition should not be drawn from isolated type 2 cognition, hence, we cannot infer underlying implicit and unconscious inferential mechanisms from behavioural manifestations of deliberate reasoning. This is an inherent contradiction in computational psychiatry, which is claiming to look at underlying implicit, automatic inferential processes while experimentally isolating overt manifestations, hence explicit reasoning. In other words, experiments are claiming to get behind the hidden structures of implicit cognition, which are assumed to be inferential, although, with the commonly used paradigms i.e., the beads task, they are targeting type-2 cognition, hence, explicit reasoning or decision-making. Even in tasks that target more perceptual processes, delusions are mostly still operationalised as unfolding not within the perceptual process, but within perceptual decision-making. This mathematical re-description of the isolated explicit form of cognition is, through formalization, then given a *law*-like normative status.

It should be kept in mind that computational approaches currently re-describe and formalize overt functional manifestations. And manifestation rarely represent the underlying route cause or core invariant feature of a phenomenon. There might be underlying inferential processes that contribute to cognition, but these would require

actual empirical prove, confirming the psychological assumption that the mind functions like a digital computer, by following heuristic unconscious operations.

If the psychological assumption cannot be proven, perhaps the epistemological assumptions (Dreyfus, 1992) applied to psychopathology, namely that people do not follow unconscious inferential operations, but their behaviour may still be descripted and formalised in terms of rules. However, the very usefulness of doing so should then be stated. Thus, if the psychological assumption can currently not be proven, we are left with a formal re-description of the performance, rather than an explanation of the very phenomenon. This is best illustrated with the example of a cyclist, that is not unconsciously following a rule to keep stable. Yet, formalization might help us to understand the competency of keeping stable on a bike. However, it does not tell us anything about what is going on in the cyclist's mind to keep stable. Similarly, computational models might tell us what it *is* to make a decision under specific circumstances, but they tell us very little of what is actually going on in a delusional person's mind. Thus, there is a huge difference between a timeless and contextless theory or description or re-description of competency and an actual theory of human behaviour or performance. In other words, to assume that humans unconsciously go through a series of Bayesian operations when making a decision or perceiving is as saying the plants are solving differential equations when making their way around the sun, even if the movement of plants can technically be described in that way. Thus, only because we can describe a process in terms of discrete elements does not mean that the process is actually achieved like that. In short, while the psychological assumptions supposes that the rules used in formalisation of behaviour and cognition are the very same rules which produce the behaviour, the epistemological assumption only affirms that all non-arbitrary behaviour can potentially be formalised according to rules. And in order to formalise a phenomena, we need to know all factors (input, goals,.etc) that lead to a phenomenon, which, due to the explanatory pluralism of psychiatric phenomena, might be an impossible endeavour. As rational analysis is not a theory of underlying psychological processes, it becomes questionable to which extent rational analysis can be a theory of underlying algorithmic processes relevant for psychopathology. Furthermore, it might thus be useful to ask whether we are testing an underlying mechanism or re-describing a manifestation. Both processes should be distinguished from one another.

Predictive coding and especially the *one-factor account* or *canonical view* (Sterzer et al., 2018) as a response to the two-factor *theory of delusions* explicitly hopes to overcome the distinction between *lower* automatic and perceptual and higher cognitive states, as all states are hierarchically arranged to make a hypothesis about the most likely cause of sensory activity (or hidden world states). Proponents of the canonical predictive coding account (Corlett et al., 2019; Sterzer et al., 2018) conclude that all abnormalities in perception and reasoning (thus two factors), can be explained with the same kind of underlying deficit in an abnormally functioning inferential system, no longer are two different deficit systems required.

This same type of deficit, in different parts of the system, can then produce a range of different phenomena. While defending the canonical view, some paradigms still located these deficits more on perceptual levels (Adams et al., 2021; Sterzer et al., 2018), while other computational paradigms proclaim inferential deficits at higher-

order prior levels, moving up the cortical hierarchy (Ashinoff et al., 2021; Baker et al., 2019). Importantly, we are not able to capture the process of making the most likely guess about the cause of sensory activity or a hidden state that unconscious inferences might make in behavioural paradigms that target explicit and deliberate reasoning. Processes involved in the psychopathology of psychosis may be inferential, automatic, and unconscious, however, current behavioural paradigms do not isolate these processes successfully.

To gauge whether the underlying algorithm simulates something that is automatic and implicit, it should be asked whether the type of cognition that is isolated is of an explicit and deliberate or implicit and automatic kind. Furthermore, even though on the computational level the same types of algorithms may be used to describe different types of cognitive processes, that does not equate these processes on the functional or experiential level, which is the level targeted in experimental tasks to collect the outputs going into the model parameters. "*Seeing is believing*" therefore applies cognition and experience, not to the computational level. Depending on the type of manifestation, being perceptual or higher-order and cognitive, it could be asked how useful an inferential model is as a hypothesis to test that very function. Thus, it also becomes questionable how computational psychiatry bridges different levels of explanation using the one-factor account; or how the one-factor account can bring together phenomenology, behavior and physiology. The one-factor account, the argument of one underlying process, has also been used to make inferences from translational accounts i.e. animal models (see Schmack et al., (2022)) to phenomenology and physiology. However, as we have seen, a shared underlying formalism of the assumed mechanism should not be used as an argument for a shared functionality and inferred shared phenomenology or experience, which is of utmost importance when designing experimental paradigms.

Computational psychiatry heavily relies on explicit rational optimality, which is then given a law-like rather than a monotonic status. However, the underlying algorithms' inferential rationality is based on functions in a similar way as in logic and are both related to a shared *truth*, hence, single objective reality. Different than in logic, probability refers to non-monotonicity, however, within most computational paradigms, this probabilistic or *statistical truth* is still misrepresented as an absolute one. Furthermore, the deontic or law-like status of computational psychiatry is rooted in assumptions about the most optimally rational solution to deal with uncertainty. Computational psychiatry therefore depends on the notion of rationality to define and gauge psychopathology, whereas violations of rationality are of an inferential kind.

Reasoning impairments are the central part of computational paradigms. Computational psychiatry, though often trying to align itself closer with biological and neurological approaches, still entails a fundamental cognitive assumption about human psychology as rational agents, and is thereby an extension to the cognitivist tradition, which becomes apparent through the normative approach. *As stated by* (Williams, 2020, p. 11), *"If one abandons the assumption (of cognitivism), that cognition is fundamentally determined by rationality and optimality, which dates back to logic (the brain as a logical or inferential machine), it becomes less clear what the initial motivation for the Bayesian brain hypothesis actually is".*

Further, rationality is directly contingent on the optimal way of dealing with uncertainty embedded in the statistical structure of the immediate environment or task, thus, a particular state of the world, which we defined as problem- or goal-dependency. The *optimal* or *ideal* way of integrating evidence in a given model is fixed and applies to every agent finding themselves under that evidence in the same way. Thus, normative models operate under the assumption of a stable single external reality that is used by multiple agents in the same way, unless the agent imposes motivational or emotional biases. In short, it is assumed that people *reason* or make inferences about the uncertainty of the world in a uniform way, a way whose ideal or optimal form can be predefined with inferential models. In terms of this dependency, when using computational approaches, theories should be built on how the input structure of the environment, creating uncertainty, as well as a universally as optimal defined way to deal with that structure are related to the aetiology of psychopathology. Fixed uncertainty deprives uncertainty itself of meaning and context, here uncertainty is of a mathematical rather than of a semantic kind. Akin to that, we need a conceptual justification why a general and rule-based reduction of uncertainty in a shared external environment plays a role in psychopathology. Reducing uncertainty may play a role in psychopathology, however, uncertainty might not arise in domain-general ways and might neither influence individuals in a unified way. In contrast to the free-energy principle, hence, the rule-based minimisation of uncertainty or prediction error, due to Dreyfus (1992), humans can deal with disambiguation in a way that uncertainty is reduced due to subjective goals and concerns in a given situation. Thus, it is the situation and the agent that determines how to disambiguate the facts. In a sharp contrast to that, to consider a context, an algorithm must either treat some features as intrinsically relevant, or it will be faced with infinite regress of contexts. Furthermore, humans naturally exhibit an ambiguity tolerance, i.e., for odd, non-rule-like behaviours. This again stands in sharp contrast to formalization that does not tolerate any form of ambiguity. The existence of other forms of rationality, such as procedural rationality, epistemic rationality and agential rationality makes it clear that rationality, in a mathematical or non-mathematical way, is a concept that is applied to behavioural manifestations, rather than to underlying implicit mechanisms. Thus, it could be asked how useful rationality is as a concept to think of underlying implicit and unconscious mechanisms.

Further, it should be asked whether an inferential deontic approach is an appropriate benchmark to define psychopathology. In terms of rationality playing a conceptual role in psychopathology, it can be questioned to which extent rationality in a philosophical sense of being *understandability* conceptually relates to optimality in a mathematical sense and if that conception is applicable to both, automatic and deliberate processes.

In terms of rationality becoming optimality, we have seen how the determination of the context is crucial for the definition of optimal. In other words, optimality is baked into the task design. Here again, we would need a theoretical justification of how a specific context, the environmental input structure, is actually relevant for psychopathology. Both, definition of the construct as well as the stipulation of the context, hence, task design, are directly built on ontological commitments. Thus, often in computational psychiatry, we see a conceptual gap between the ontological com-

mitments and what is actually being tested in the experimental paradigms: Processes within the BBH do not refer to deliberate decisions but to a strategy of free-energy minimisation, that has very little to do with deliberate reasoning. However, this very concern is increasingly addressed with paradigms bypassing deliberate reasoning, as seen in Corlett's group and others.

Campbell, (2001) questioned the direct transformation from evidence to a propositional framework leading to delusions. Rationality, regardless of it being logical or probabilistic rationality, appeals to a shared propositional framework, which is of a mathematical kind and entails predetermined goals. However, we would need a justification why the particular goals and universally predetermined limitations of the agent or the predetermined constraints of the environment are relevant for psychopathology. Internalized goals may play a role in psychopathology however, these goals might be personal, rather than presenting a generic mathematical reduction of fixed input uncertainty. *Similarly, to apply inferential laws to psychopathology, Quantum psychopathology (QPP) applies laws and generalisations from quantum physics to the brain and psychological processes. Thereby, QPP may reach beneath and beyond the boundaries of discrete neurons and Newton mechanistic (Tarlacı, 2019; Malik and Lindesay, 2009), hence, the inferential logic of computational psychiatry, while QPP still also represents a reductionist approach. The inferential deterministic model, in contrast to QPP, aims at the prediction of a later state, given that all parameters describing the previous state in the system are known (Tarlacı, 2019). In this paper, we have seen that this is an extensive problem for the computational modelling of complex states such as delusions, as we do not know all factors that determine the state. QPP goes on step further and challenges the general notion of cause-and-effect models and deterministic predictions in general, while suggesting models from quantum mechanics. The hope of QPP is that new analogies between quantum physics and psychiatric illnesses could help to understand the later. Importantly and in contrast to computational psychiatry, QPP highlights the nature of this analogy and that theories might not directly be testable (Tarlacı, 2019; Malik and Lindesay, 2009). One application of quantum mechanics to psychopathology is Niels Bohr's multiple or parallel universe model , which questions our understanding of a single reality. Psychiatric diagnosis as well as computational psychiatry relies on the assumption of a single objective reality in which psychopathology unfolds and can be measured against an (inferential) 'norm'. The entire notion of 'rationality', in either a logical or Bayesian way, heavily relies on the premise of a single objective reality. However, as suggested in this paper, the objectively shared ontological reality might not be the realm of psychopathology, an argument that can further be backed up by the new insights and epistemological process into the fundamental nature of reality, hence, the multiple universe model of quantum mechanics. In that vein, input-output determinism has been superseded by quantum physics. Furthermore, it is important to mention that computational psychiatry as well as quantum psychopathology both represent attempts to apply theoretical models from other disciplines, such as mathematics and physics, to psychopathology, which represents a theory-driven approach unfolding in the respective paradigms. However, as highlighted in the paper, a full theory-driven analogy might sometimes be useful but can also become misleading and inappropriate. Further, it represents one way, next to more data-driven (e.g.,*

phenomenologically data driven) approaches, that highlight individual meaning detached from underlying theory.

In a similar vein, it remains questionable whether Marr's levels of explanation can be meaningfully related to psychopathology, especially to the direct input-output dependency. Using Marr's levels of computation implies that psychopathology is a direct result of a given input-output mapping and therefore also of the immediate environmental structure feeding into the input. Thus, it is assumed that the chosen type of input is directly relevant and translates into psychopathology.

In other words, dependency on the input directly implies that the concrete input of an experiment elicits the phenomena in questions, hence the delusions, in a domain general way. Therefore, computational models should be interpreted carefully in terms of what the model is actually isolating in terms of input and output, and how all of these factors are relate to psychopathology.

In terms of theory-building, we can only formalize what we can sufficiently define, which highlights the need for valid, robust, and distinct concepts. This fact alone calls the formalization of psychopathological phenomena into question, whose definitions are often based on conventions, rather than robust natural phenomena. This directly translated into the validity of the operationalisation and measurement technique; a task might measure something like *believing external advice against evidence*, however, the relevance of this very cognition for psychopathology needs to be stipulated.

It becomes apparent that concepts are often operationalised in a way that makes them formalizable.

As (Maatman, 2021) stated, formalization alone does not identify underlying mechanisms, these need to be specified beforehand, the elements of the experiment, such as the definition or the ontological commitments, will not become apparent nor specific due to modelling alone. Neither can we identify causal connections between variables on the basis of their mathematical relationship. Formal modelling is not suited for phenomena that are too complex, thereby forcing the researcher into simplified operationalisations and designs (Maatman, 2021).

Our critique of the assumptions of computational psychiatry can be integrated with the Kuhnian assumption of science being in flux, where the computational approach to science, including the mind and psychopathology has become the conventional basis and received view for research for the time being and failures to empirical demonstrate theories currently do not cause the rejection of the theory and background assumptions themselves. The Quine-Duhem thesis and especially holistic underdetermination (Maatman, 2021), illustrate the web of background theories and assumptions empirical computational psychiatry is embedded in. As highlighted in the paper, some being background assumptions about the computational and specially the inferential mode of operation of the mind, rationality as guiding principle and cognitivism. A failed hypothesis typically leaves open the possibility of abandoning one of these background beliefs and assumptions rather than the hypothesis itself, which is, however, barely done in computational psychiatry. Thus, not questioning the background hypotheses, the status of computationalism as the received view remains. This could also be described as a disconnect and an underdetermination of alternative paradigms or ontological models of psychopathology, i.e., concerning the human existence and experience in more humanistic disciplines, such as phenom-

enology, first-person accounts or enactivism. These accounts or paradigms can currently not be integrated into the web of computational theories and ontologies, and as the result of the immensurability of competing paradigms, described by Kuhniamism, are often rendering unscientific or are not taken seriously. Furthermore, they belong to a fundamentally different set/web of theories, which are not grounded in the inferential and computational conception of the mind and input-output determinism. Thus, within conceptual underdetermination, the same experimental data might be explainable with entirely different paradigms, alternatives to computational ones, which are, however, currently not the dominant view. Therefore, more humanistic vs. computational approaches represent two contrasting Kuhnian paradigms, with entirely different ontological assumptions (Broeker and Broome, in submission).

There are arguments demonstrating that any judgement can be modelled in terms of Bayesian inference by making suitable assumptions about priors and likelihoods, which does not tell us that it is a useful approach to use. If a computational model is applied to psychopathology, it should be clearly stated how explicit inferential mechanisms apply to theories of psychopathology. Computational psychiatry often fails to notice that human suffering is meaningful and not mechanistically caused by an inferential failure mode, where any intrinsic relationship between the components is missing. Computational psychiatry has become so focussed on the form of the symptoms, that it fails to notice that symptoms are meaningful and a form of communication.

Acknowledgements We would like to thank Dr. Brandon Ashinoff, Dr. Clara Humpston and Dr. Phil Corlett for useful conversations and comments on previous versions of the draft.

Declarations

Conflict of interest None reported.

References

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2015). Computational Psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology Neurosurgery*. https://doi.org/10.1136/jnnp-2015-310737. Psychiatryjnnp-2015-310737.

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*, 47–47. https://doi.org/10.3389/fpsyt.2013.00047

Adams, R. A., Vincent, P., Benrimoh, D., Friston, K. J., & Parr, T. (2021). Everything is connected: inference and attractors in delusions. *Schizophrenia Research*, 0920996421003054. https://doi.org/10.1016/j.schres.2021.07.032

Agich, G. J. (2002). Implications of a pragmatic theory of disease for the DSMs. *Descriptions and prescriptions: values, mental disorders, and the DSMs., ed. J. Z. Sadler* (pp. 96–113). The Johns Hopkins University Press.

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates.

Apa (1987). *Diagnostic and Statistical Manual of Mental Disorders (3rd ed., revised)*.

Apa. (2013). *Dsm 5 Diagnostic and Statistical Manual of Mental Disorders 5 Ed Spl Edition*. Cbs Publishing.

Ashinoff, B. K., Singletary, N. M., Baker, S. C., & Horga, G. (2021). Rethinking delusions: a selective review of delusion research through a computational lens. *Schizophrenia Research*, 0920996421000657. https://doi.org/10.1016/j.schres.2021.01.023

Baker, S. C., Konova, A. B., Daw, N. D., & Horga, G. (2019). A distinct inferential mechanism for delusions in schizophrenia. *Brain (London England:, 1878)*, *142*(6), 1797–1812. https://doi.org/10.1093/brain/awz051

Bansal, S., Bae, G. Y., Robinson, B. M., Hahn, B., Waltz, J., Erickson, M., Leptourgos, P., Corlett, P., Luck, S. J., & Gold, J. M. (2022). Association between failures in Perceptual updating and the severity of psychosis in Schizophrenia. *JAMA Psychiatry*, *79*(2), 169–177. https://doi.org/10.1001/jamapsychiatry.2021.3482

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions Biological Sciences*, *371*(1708), 20160011. https://doi.org/10.1098/rstb.2016.0011

Bortolotti, L., & Broome, M. R. (2008). Delusional beliefs and reason giving. *Philosophical Psychology*, *21*(6), 821–841. https://doi.org/10.1080/09515080802516212

Broome, Matthew. (2007). Taxonomy and Ontology in Psychiatry: A Survey of Recent Literature. *Philosophy, Psychiatry, & Psychology*, *13*(4), 303–319. https://doi.org/10.1353/ppp.2007.0026

Campbell, J. (2001). Rationality, meaning, and the analysis of delusion. *Philosophy Psychiatry & Psychology*, *8*(2), 89–100. https://doi.org/10.1353/ppp.2001.0004

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Coltheart, M., Menzies, P., & Sutton, J. (2010). Abductive inference and delusional belief. *Cognitive Neuropsychiatry*, *15*(1–3), 261–287. https://doi.org/10.1080/13546800903439120

Cooper, R. (2004). What is wrong with the DSM? *History of Psychiatry*, *15*(1), 5–25. https://doi.org/10.1177/0957154X04039343

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R. (2019). Hallucinations and strong priors. *Trends in Cognitive Sciences*, *23*(2), 114–127. https://doi.org/10.1016/j.tics.2018.12.001

Cosgrove, L. (2011). The DSM, big pharma, and clinical practice guidelines: protecting patient autonomy and informed consent. *IJFAB: International Journal of Feminist Approaches to Bioethics*, *4*(1), 11–25.

Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason ([3rd ed])*. MIT Press.

ELLIS, A. (1957). Rational psychotherapy and individual psychology. *Journal of Individual Psychology*, *13*(1), 38.

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in psychology: how to move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Fahlman, S. E., Hinton, G. E., & Sejnowski, T. J. (1983). Massively parallel architectures for AI: Netl, thistle, and boltzmann machines. *Proceedings of the Third AAAI Conference on Artificial Intelligence*, 109–113.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology Paris*, *100*(1–3), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Gadsby, S., & Hohwy, J. (2019). *Why use predictive processing to explain psychopathology? The case of anorexia nervosa*. PsyArXiv. https://doi.org/10.31234/osf.io/y46z5

Ghaemi, S. N. (2009). Nosologomania:DSM &Karl Jaspers' Critique of Kraepelin. *Philosophy Ethics, and Humanities in Medicine*, *4*(1), 10.

Gigerenzer, G. (2008). *Rationality for mortals: how people cope with uncertainty*. Oxford University Press.

Griffin, J. D., & Fletcher, P. C. (2017). Predictive Processing, source monitoring, and psychosis. *Annual Review of Clinical Psychology*, *13*(1), 265–289. https://doi.org/10.1146/annurev-clinpsy-032816-045145

Haslam, N. (2002). Kinds of kinds: a conceptual taxonomy of Psychiatric categories. *Philosophy Psychiatry & Psychology*, *9*, 203–217. https://doi.org/10.1353/ppp.2003.0043

von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*. Leopold Voss. https://hdl.handle.net/2027/hvd.32044106192305

von Helmholtz, H. (Ed.). (1925). *Helmholtz's treatise on physiological optics, translated from the 3d German ed. Edited by James P.C. Southall*. The Optical Society of America-1925. https://hdl.handle.net/2027/mdp.39015067341399

Hinton, G. E., & Zemel, R. S. (1993). Autoencoders, minimum description length and Helmholtz free energy. *Proceedings of the 6th International Conference on Neural Information Processing Systems*, 3–10.

Hohwy, J. (2013). *The predictive mind*. University Press.

Horwitz, A. V. (2002). *Creating mental illness*. University of Chicago Press.

Huys, Q. J. M., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Networks*, *24*(6), 544–551. https://doi.org/10.1016/j.neunet.2011.03.001

Jardri, R., & Denève, S. (2013). Circular inferences in schizophrenia. *Brain: A Journal of Neurology*, *136*(Pt 11), 3227–3241. https://doi.org/10.1093/brain/awt257

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and Biases*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809477

Kendell, R., & Jablensky, A. (2003). Distinguishing between the Validity and Utility of Psychiatric Diagnoses. *American Journal of Psychiatry*, *160*(1), 4–12. https://doi.org/10.1176/appi.ajp.160.1.4

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, *8*, 302–302. https://doi.org/10.3389/fnhum.2014.00302

Maatman, F. O. (2021). *Psychology's Theory Crisis, and Why Formal Modelling Cannot Solve It*. PsyArXiv. https://doi.org/10.31234/osf.io/puqvs

Malik, M. A., & Lindesay, J. (2009). Quantum Physics: Relevance to Psychiatry. NeuroQuantology, 7(2).

Manktelow, K. I., & Over, D. E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition*, *39*(2), 85–105. https://doi.org/10.1016/0010-0277(91)90039-7

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. WHFreeman.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39–39. https://doi.org/10.3389/fnhum.2011.00039

McLean, B. F., Mattiske, J. K., & Balzan, R. P. (2017). Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: a detailed Meta-analysis. *Schizophrenia Bulletin*, *43*(2), 344–354. https://doi.org/10.1093/schbul/sbw056

Mishara, A. L. (2007). Missing links in phenomenological clinical neuroscience: why we still are not there yet. *Current Opinion in Psychiatry*, *20*(6), 559–569. https://doi.org/10.1097/YCO.0b013e3282f128b8

Mishara, A. L., & Sterzer, P. (2015). Phenomenology *is* bayesian in its application to delusions. *World Psychiatry*, *14*(2), 185–186. https://doi.org/10.1002/wps.20213

Miyazono, K., & Bortolotti, L. (2021). *Philosophy of Psychology: An Introduction* (1st edition). Polity.

Oaksford, M., & Chater, N. (2009). Précis of *bayesian rationality: the Probabilistic Approach to Human reasoning*. *Behavioral and Brain Sciences*, *32*(1), 69–84. https://doi.org/10.1017/S0140525X09000284

O'Callaghan, C., Hall, J. M., Tomassini, A., Muller, A. J., Walpola, I. C., Moustafa, A. A., Shine, J. M., & Lewis, S. J. (2017). Visual hallucinations are characterized by impaired sensory evidence Accumulation: insights from hierarchical drift diffusion modeling in Parkinson's Disease. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *2*(8), 680–688. https://doi.org/10.1016/j.bpsc.2017.04.007

Parnas, J., & Zahavi, D. (2002). The role of phenomenology in psychiatric diagnosis and classification. *Psychiatric diagnosis and classification* (pp. 137–162). John Wiley & Sons Inc. https://doi.org/10.1002/047084647X.ch6

Piaget, J. (1957). Études d'épistémologie génétique. *Études d'épistémologie génétique*. Presses universitaires de France.

Pickersgill, M. D. (2014). Debating DSM-5: diagnosis and the sociology of critique. *Journal of Medical, & Ethics*, *40*(8), 521–525.

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction Vulnerabilities in the decision process. *The Behavioral and Brain Sciences*, *31*(4), 415–437. https://doi.org/10.1017/S0140525X0800472X

Reed, E. J., Uddenberg, S., Suthaharan, P., Mathys, C. D., Taylor, J. R., Groman, S. M., & Corlett, P. R. (2020). Paranoia as a deficit in non-social belief updating. *ELife*, *9*. https://doi.org/10.7554/eLife.56345

Ritunnano, R., Kleinman, J., Oshodi, D. W., Michail, M., Nelson, B., Humpston, C. S., & Broome, M. R. (2022). Subjective experience and meaning of delusions in psychosis: a systematic review and qualitative evidence synthesis. *The Lancet Psychiatry*, *0*(0), https://doi.org/10.1016/S2215-0366(22)00104-3

Rossi-Goldthorpe, R. A., Leong, Y. C., Leptourgos, P., & Corlett, P. R. (2021). Paranoia, self-deception and overconfidence. *PLOS Computational Biology*, *17*(10), e1009453. https://doi.org/10.1371/journal.pcbi.1009453

Sadler, J. Z. (2004). Diagnosis/antidiagnosis. In *The philosophy of psychiatry: A companion* (pp. 163–179). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195149531.001.0001

Samuels, R., Stich, S., & Bishop, M. (2002). Ending the Rationality Wars: How To Make Disputes About Human Rationality Disappear. *Institute of Philosophy*. https://sas-space.sas.ac.uk/938/

Schmack, K., Ott, T., & Kepecs, A. (2022). Computational Psychiatry Across Species to Study the Biology of Hallucinations. *JAMA Psychiatry*, *79*(1), 75–76. https://doi.org/10.1001/jamapsychiatry.2021.3200

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., & Friston, K. (2016). Corrigendum to 'Optimal inference with suboptimal models: Addiction and active Bayesian inference' [Med. Hypotheses 84 (2015) 109–117]. *Medical Hypotheses*, *91*, 123–123. https://doi.org/10.1016/j.mehy.2016.02.021

Shea, N., & Frith, C. D. (2016). Dual-process theories and consciousness: the case for "Type Zero" cognition. *Neuroscience of Consciousness*, *2016*(1), niw005–niw005. https://doi.org/10.1093/nc/niw005

Skene, A. (2002). Rethinking normativism in psychiatric classification. *In descriptions and prescriptions: values, mental disorders, and the DSMs, ed. J. Z. Sadler* (pp. 114–127). The Johns Hopkins University Press.

Stanovich, K. (1999). *Who Is Rational? Studies of Individual Differences in Reasoning*.

Stephens, G. L., & Graham, G. (2004). Reconceiving delusion. *International Review of Psychiatry (Abingdon England)*, *16*(3), 236–241. https://doi.org/10.1080/09540260400003982

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, *84*(9), 634–643. https://doi.org/10.1016/j.biopsych.2018.05.015

Sullivan-Bissett, E., & Noordhof, P. (2020). The transparent failure of norms to keep up Standards of Belief. *Philosophical Studies*, *177*(5), 1213–1227. https://doi.org/10.1007/s11098-019-01242-y

Suthaharan, P., Reed, E. J., Leptourgos, P., Kenney, J. G., Uddenberg, S., Mathys, C. D., Litman, L., Robinson, J., Moss, A. J., Taylor, J. R., Groman, S. M., & Corlett, P. R. (2021). Paranoia and belief updating during the COVID-19 crisis. *Nature Human Behaviour*, *5*(9), 1190–1202. https://doi.org/10.1038/s41562-021-01176-8

Tarlacı, S. (2019). Quantum neurobiological view to mental health problems and biological psychiatry. *Journal of Psychopathology*.

Thornton, T. (2002). *Reliability and validity in psychiatric classification: values and neo-humeanism*. Philosophy, Psychiatry, & Psychology.

Tripoli, G., Quattrone, D., Ferraro, L., Gayer-Anderson, C., Rodriguez, V., La Cascia, C., La Barbera, D., Sartorio, C., Seminerio, F., Tarricone, I., Berardi, D., Szöke, A., Arango, C., Tortelli, A., Llorca, P. M., de Haan, L., Velthorst, E., Bobes, J., Bernardo, M., & Di Forti, M. (2020). *Jumping to conclusions, general intelligence, and psychosis liability: Findings from the multi-centre EU-GEI case-control study*. https://discovery.ucl.ac.uk/id/eprint/10096697

Weizenbaum, J. (1976). *Computer power and human reason: from judgment to calculation*. WHFreeman.

Williams, D. (2020). Epistemic irrationality in the bayesian brain. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz044

Williams, D., & Montagnese, M. (2020). *Bayesian Psychiatry and the Social Focus of Delusions*. https://doi.org/10.13140/RG.2.2.27852.23683

Wittgenstein, L. (1974). *On certainty*. Blackwell.

Authors and Affiliations

Marianne D. Broeker[1] · Matthew R. Broome[2]

✉ Marianne D. Broeker
marianne.broeker@psy.ox.ac.uk

Matthew R. Broome
M.R.Broome@bham.ac.uk

[1] Department of Experimental Psychology, University of Oxford, St. Anne's College, 56 Woodstock Rd, OX2 6HS Oxford, UK

[2] Institute for Mental Health, University of Birmingham, Birmingham, UK