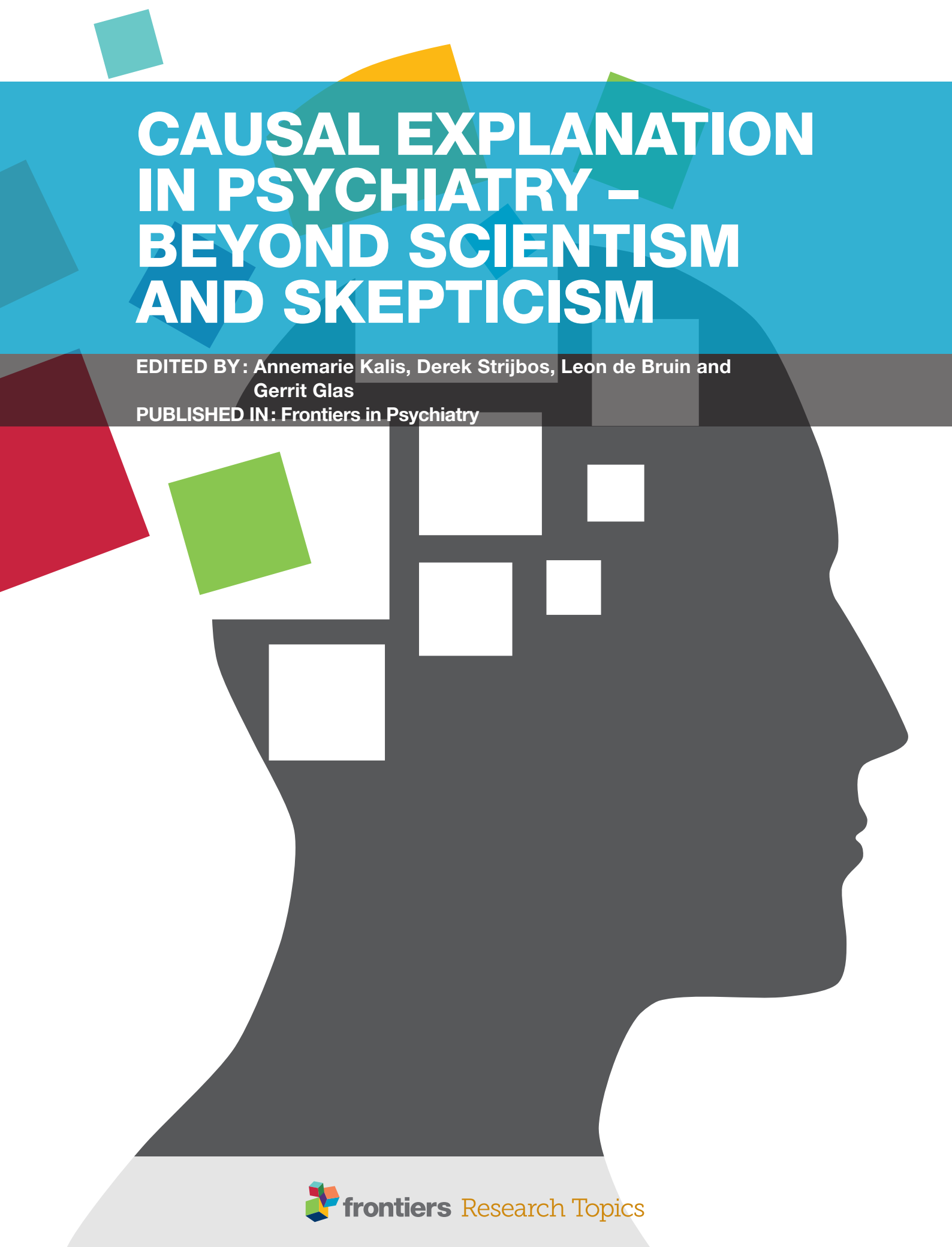
CAUSAL EXPLANATION IN PSYCHIATRY – BEYOND SCIENTISM AND SKEPTICISM

EDITED BY: Annemarie Kalis, Derek Strijbos, Leon de Bruin and Gerrit Glas

PUBLISHED IN: Frontiers in Psychiatry



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-229-3
DOI 10.3389/978-2-88945-229-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

CAUSAL EXPLANATION IN PSYCHIATRY – BEYOND SCIENTISM AND SKEPTICISM

Topic Editors:
Annemarie Kalis, Utrecht University, Netherlands
Derek Strijbos, Radboud Universiteit Nijmegen, Netherlands
Leon de Bruin, Radboud Universiteit Nijmegen, Netherlands
Gerrit Glas, VU University, Netherlands

Citation: Kalis, A., Strijbos, D., de Bruin, L., Glas, G., eds. (2017). Causal Explanation in Psychiatry – Beyond Scientism and Skepticism. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-229-3

Table of Contents

frontiers in Psychiatry

EDITORIAL
published: 09 May 2017
doi: 10.3389/fpsyt.2017.00070

Editorial: Causation and Causal Explanation in Psychiatry—Beyond Scientism and Skepticism

Annemarie Kalis[1], Derek Strijbos[2], Leon de Bruin[3] and Gerrit Glas[3]*

[1] Utrecht University, Utrecht, Netherlands, [2] Radboud University, Nijmegen, Netherlands, [3] VU University, Amsterdam, Netherlands

Keywords: psychiatry, causal processes, integration, complexity, mental states

The Editorial on the Research Topic

Causation and Causal Explanation in Psychiatry—Beyond Scientism and Skepticism

Since psychiatry firmly established itself as a scientific discipline, it has been propelled forward by the hope that the different diagnostic categories distinguished in clinical practice, will turn out to correspond to unique underlying causes. However, so far there is little evidence that disorders such as major depression or schizophrenia can be traced back to relatively simple, common causal trajectories. Rather, the etiology of almost all mental disorders seems to be complex and multifactorial and to span different levels of explanation, ranging from (epi)genetic, neurobiological to psychological, and social levels.

Clinicians, broadly speaking, tend to be skeptical about the prospects of causal modeling in psychiatry, whereas scientists tend to cling to a scientistic and sometimes also reductionistic view on mental disorder. Psychiatry needs to find a way beyond skepticism and scientism, and this requires new methods and new conceptual approaches that enable us to gain a better insight into the complexity of the causal processes leading to mental disorders.

This Research Topic discusses novel theoretical and empirical strategies addressing causation and causal explanation in psychiatry, in the context of a broader discussion of what science can and cannot contribute to the definition of mental disorder. Questions addressed are: how could the complexity of mental disorders be modeled and empirically investigated? Are traditional nomological theories of causation the best framework for thinking about causation in psychiatry, or should we look at alternatives such as mechanism-based, interventionist, or pluralist theories of causation? How to integrate different levels of explanation in etiological models of mental disorder?

Dijkstra and de Bruin investigate to what extent it is justified to draw conclusions about causal relations between brain states and mental states from "traditional" cognitive neuroscience studies and brain stimulation studies. They argue that, depending on whether one adopts Woodward's or Baumgartner's interventionist account of causation, it is possible to draw causal conclusions from both types of studies (Woodward) or from brain stimulation studies only (Baumgartner). Also, they show what happens to these conclusions if we adopt different views of the relation between mental states and brain states.

Gijsbers reviews recent debates about the unity of science and explanatory pluralism, focusing on the tension between the integrative and the isolationist perspective: should the integrative tendencies in science be fully indulged in, or is a certain amount of isolation necessary? He argues that an important question is whether two true explanations of the same fact can ever fail to be combinable into one single explanation and shows that this can be the case when explanations have incompatible counterfactual consequences. He thus concludes that although interdisciplinarity may have many advantages, we should not take the project of integration too far.

OPEN ACCESS

Edited and Reviewed by:
Raina Robeva, Sweet Briar College, USA

****Correspondence:***
Annemarie Kalis
a.kalis@uu.nl

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry

Received: *27 February 2017*
Accepted: *13 April 2017*
Published: *09 May 2017*

Citation:
Kalis A, Strijbos D, de Bruin L and Glas G (2017) Editorial: Causation and Causal Explanation in Psychiatry—Beyond Scientism and Skepticism. Front. Psychiatry 8:70. doi: 10.3389/fpsyt.2017.00070

According to Hutto, philosophy of psychiatry faces a tough choice between two competing ways of understanding mental disorders. The folk psychology (FP) view puts our everyday normative conceptual scheme in the driver's seat. Opposing this, the scientific image (SI) view holds that our understanding of mental disorders must come from the mind sciences. This paper rejects both the FP view (in its pure form) and the SI view, in its popular cognitivist renderings. It concludes that a more liberal version of SI can accommodate what is best in both views and provide a sound philosophical basis for a future psychiatry.

Thornton focuses on the idea that psychiatry contains, in principle, a series of levels of explanation—an idea that has been criticized as presupposing a discredited pre-Humean view of causation. These claims echo some superficially similar remarks in Wittgenstein's Zettel. Thornton argues that attention to the context of Wittgenstein's remarks suggests a reason to reject explanatory minimalism in psychiatry and reinstate a Wittgensteinian notion of levels of explanation.

Van Riel starts from the common assumption that social environment and cultural formation shape mental disorders. The details of this claim are, however, not well understood. His paper takes a look at the claim that culture has an impact on psychiatry from the perspective of metaphysics and the philosophy of science. Its aim is to offer, in a general fashion, partial explications of some significant versions of the thesis that culture and social environment shape mental disorders and to highlight some of the consequences social constructionism about psychiatry has for psychiatric explanation.

Stein and Illes discuss the emergent field of global mental health, which has paid particular attention to upstream causal factors, for example, poverty, inequality, and gender discrimination in the pathogenesis of mental disorders. However, this field has also been criticized for relying erroneously on Western paradigms of mental illness. The authors argue that it is important to steer a path between scientism (disorders as essential categories) and skepticism (disorders as mere social constructions) and propose an integrative model that emphasizes the contribution of a broad range of causal mechanisms and the consequent importance of broad spectrum approaches to intervention.

Young presents a hybrid top-down, bottom-up model of the relationship between symptoms and mental disorder, viewing symptom expression and their causal complex as a reciprocally dynamic system with multiple levels, from lower-order symptoms in interaction to higher-order constructs affecting them. He concludes that symptoms vary over several dimensions, including: subjectivity, objectivity, conscious motivation effort, and unconscious influences, and discusses the degree to which individual (e.g., meaning) and universal (e.g., causal) processes are involved.

Bechtel reviews some of the compelling evidence of disrupted circadian rhythms in individuals with mood disorders (major depressive disorder, seasonal affective disorder, and bipolar disorder). While the evidence is suggestive of an etiological role for altered circadian rhythms in mood disorders, it is compatible with other explanations. In light of this, the paper advances a proposal as to what evidence would be needed to establish a direct causal link between disruption of circadian rhythms and mood disorders.

Bielczyk et al. integrate the literature on cognitive and physiological biomarkers of MDD with the insights derived from mathematical models of brain networks. They propose a new approach called "circuit to construct mapping," which aims to characterize causal relations between the underlying network dynamics (as the cause) and the constructs referring to the clinical symptoms of MDD (as the effect).

AUTHOR CONTRIBUTIONS

AK, DS, LB, and GG wrote and approved the manuscript.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

HYPOTHESIS AND THEORY
published: 19 July 2016
doi: 10.3389/fpsyt.2016.00129

Cognitive Neuroscience and Causal Inference: Implications for Psychiatry

Nadine Dijkstra[1] and Leon de Bruin[2]*

[1] Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, Nijmegen, Netherlands, [2] Department of Philosophy, VU University Amsterdam, Amsterdam, Netherlands

In this paper, we investigate to what extent it is justified to draw conclusions about causal relations between brain states and mental states from cognitive neuroscience studies. We first explain the views of two prominent proponents of the interventionist account of causation: Woodward and Baumgartner. We then discuss the implications of their views in the context of traditional cognitive neuroscience studies in which the effect of changes in mental state on changes in brain states is investigated. After this, we turn to brain stimulation studies in which brain states are manipulated to investigate the effects on mental states. We argue that, depending on whether one sides with Woodward or Baumgartner, it is possible to draw causal conclusions from both types of studies (Woodward) or from brain stimulation studies only (Baumgartner). We show what happens to these conclusions if we adopt different views of the relation between mental states and brain states. Finally, we discuss the implications of our findings for psychiatry and the treatment of psychiatric disorders.

Keywords: interventionism, causal exclusion problem, cognitive neuroscience, psychiatry, mental causation

OPEN ACCESS

Edited by:
Firas H. Kobeissy,
University of Florida, USA

Reviewed by:
David Papo,
Technical University of Madrid, Spain
Eleftheria Pervolaraki,
University of Leeds, UK
Ying Xu,
University at Buffalo, USA

****Correspondence:***
Nadine Dijkstra
n.dijkstra@donders.ru.nl

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry

Received: *08 December 2015*
Accepted: *07 July 2016*
Published: *19 July 2016*

Citation:
Dijkstra N and de Bruin L (2016) Cognitive Neuroscience and Causal Inference: Implications for Psychiatry. Front. Psychiatry 7:129. doi: 10.3389/fpsyt.2016.00129

INTRODUCTION

Traditionally, cognitive neuroscientists have been probing the relation between brain states and mental states by manipulating the mental state of the participant through different conditions and then measuring the associated changes in neural activity, for example by means of Functional Magnetic Resonance Imaging (fMRI) or Electro-encephalogram (EEG). The results of these manipulations are usually taken to reflect a *correlation* between mental states and brain states, rather than a "genuine" *causal* relation. According to several neuroscientists, however, new brain stimulation techniques, such as Deep Brain Stimulation (DBS) and Transcranial Magnetic Stimulation (TMS), allow us to go beyond correlations and establish causal relations between mental states and brain states [for a review, see Ref. (1)]. This has important implications for other disciplines in which these techniques become increasingly popular. For example, in psychiatry, DBS has proven to be an effective treatment for patients with major depressive disorder (MDD) who do not respond to pharmacotherapy or psychotherapy (2–4).

In the current paper, we investigate whether and to what extent it is indeed justified to draw conclusions about causal relations between brain and mental states on the basis of cognitive neuroscience studies. In the next section, we start with a description of an interventionist account of causation, which is inspired by Woodward (5). We argue that this account is more or less in line with how causation is understood in scientific practice. The question is, however, whether it can be used to make causal claims about the interaction between mental states and brain states. In order to address this question, we introduce the notion of supervenience in Section "Mental States and Brain States: A Supervenience Relation." This notion aims to capture the intuition that mental states

are dependent on, but not identical with, brain states. In Section "Causation in Traditional Cognitive Neuroscience Studies," we turn to Baumgartner's "causal exclusion" argument. According to this argument, the assumption of a supervenience relation violates the criteria of what counts as a good intervention. As a result, we cannot draw conclusions about the causal relation between mental states and brain states. In his reply to Baumgartner, Woodward (6) proposes to adjust these intervention criteria in order to make room for supervenience relations and to secure causal claims on the basis of traditional cognitive neuroscience studies. In Section "Causation in Brain Stimulation Studies," we discuss the consequences of both positions for causal claims on the basis of brain stimulation studies. Most importantly, we will show that Baumgartner's causal exclusion argument does not apply to these studies. That is, we can make causal claims about brain stimulation studies *even* if we assume a supervenience relation and accept Woodward's original intervention criteria. In Section "Articulating the Mind–Brain Relation," we show what happens to these conclusions if we adopt a different view of the relation between mental states and brain states. Finally, in Section "Conclusion," we briefly discuss the implications of our findings for psychiatry and the treatment of psychiatric disorders.

THE INTERVENTIONIST ACCOUNT OF CAUSATION

In most textbooks on experimental research two main requirements are described that an experiment must meet to be able to reveal a causal relation between X and Y. The first is that the levels of X must be systematically varied and the second is that all variables other than X and Y are to be controlled in order to eliminate other possible causes of Y. If these requirements are met and changes in X are accompanied by changes in Y, one is allowed to speak of a causal relation between X and Y (7, 8).

This notion of how to investigate causal relations in scientific practice is very much in line with a philosophical account of causation that has become quite popular recently: interventionism. One of the most established interventionist definitions of causation comes from Woodward (5):

> (M) A necessary and sufficient condition for X to be a (type-level) *direct cause* of Y with respect to a variable set **V** is that there be a possible intervention on X that will change Y or the probability of Y when one holds fixed at some value all other variables Z_i in **V**. A necessary and sufficient condition for X to be a (type-level) *contributing cause* of Y with respect to variable set **V** is that (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship… and that (ii) there be some intervention on X that will change Y when all other variables in **V** that are not on this path are held fixed [Ref. (5), pp. 59].

We mainly focus on the definition of a direct cause since this comes closest to the notion of causation as it is investigated in scientific practice (i.e., it explicitly involves the two requirements mentioned above). However, for the definition to make sense, we also need a clear notion of what an appropriate intervention is. Woodward (5) defines an intervention variable as follows:

(IV) I is an intervention variable for X with respect to Y if:

1. I causes X;
2. I acts as a switch for all the other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I;
3. Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the I–X–Y connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.
4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X [(5), pp. 98].

Finally, relative to the notion of an intervention variable an (actual) *intervention* can be straightforwardly understood in terms of an intervention variable I for X with respect to Y taking on some value z_i such that $I = z_i$ causes X to take on some determinate value z_j [(5), pp. 98]. In terms of experimental design, an intervention can be seen as a manipulation that changes the variable X. In order for this manipulation to be able to reveal a causal relation, it has to meet the requirements in (IV).

MENTAL STATES AND BRAIN STATES: A SUPERVENIENCE RELATION

Can we use interventionism to make causal claims about the interaction between mental states and brain states? To answer this question, we will (initially) assume a very minimal relation between mental states and brain states – one that captures the intuition that mental states are dependent on brain states. In the philosophy of mind, this relation is known as "supervenience."

A schematic representation of a supervenience relation between mental states M1 and M2 and brain states P1 and P2 is depicted in **Figure 1**. Although the notion of supervenience has been much discussed, there are two features that are common in most definitions:

(S1) $\neg(M \text{ causes } P) \wedge \neg(P \text{ causes } M)$;
(S2) Every change in the value of M is necessarily accompanied by a change in the value of P.

This means that (i) supervenience is a non-causal relation such that neither M causes P nor vice versa[1] and (ii) any change in mental state is necessarily accompanied by a change in brain state. Furthermore, with regard to **Figure 1**, we will assume that:

(S3) P1 causes P2.

[1]Supervenience is non-causal because it represents a synchronic rather than a diachronic relation between M and P.

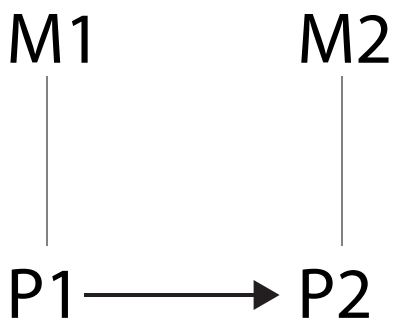


FIGURE 1 | Schematic representation of the relation between brain states and mental states. Undirected edges indicate supervenience relations and the arrow indicates a causal relation.

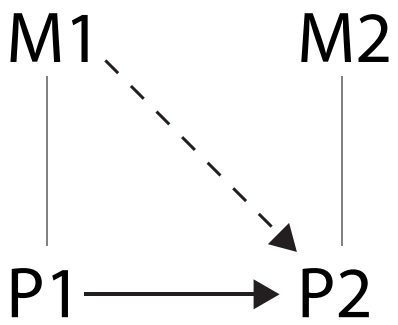


FIGURE 2 | Schematic representation of the relation investigated in traditional cognitive neuroscience studies indicated by the dashed arrow.

The end result is a schematic representation of two types of relations: one between *properties* (M1 and P1, M2 and P2), which is captured by a supervenience relation, and one between *events* (M1/P1 and M2/P2), which is captured by a causal relation (i.e., event 1 causes event 2).

CAUSATION IN TRADITIONAL COGNITIVE NEUROSCIENCE STUDIES

With the interventionist account of causation and the notion of supervenience in place, let us now take a closer look at traditional (non-invasive) cognitive science studies.

In most of these studies, the relation between mental states and brain states is investigated by observing the effect of changes in mental state M1 on brain state P2 (see **Figure 2**). This is done by manipulating the mental state of the subjects by letting them participate in separate conditions that differ on some stimulus characteristic or task that is meant to induce changes in M1. To investigate the effect of these manipulations on brain states, the subjects' brain activity P2 is measured in all conditions. Then, if the researcher has made sure that the conditions only differ on the manipulated mental variable (using all kinds of controls like randomization of subjects), and a (significant) difference in brain activity between the conditions is found, the researcher concludes that the manipulated mental variable M1 has had an effect on the measured brain state P2.

However, is it valid to conclude that the change in mental state M1 *caused* the change in brain state P2? According to the causal exclusion argument put forward by Baumgartner (9), it is not.

Baumgartner's Causal Exclusion Argument

In his argument, Baumgartner (9) takes together the interventionist definition of causation as described above in (M) and (IV) and the supervenience relation as described in (S1–2) to formulate the following conditional:

> (BM) If *M1* is causally relevant to *P2* with respect to the variable set $\mathbf{V} = \{M1, M2, P1, P2\}$, then there possibly exists a variable I_1 that causes a change in the value (or the probability distribution) of *M1* and is statistically independent of any variable *Z* that causes *P2* and that is on a directed path that does not go through *M1* [(9), pp. 170].

Now we can see that no such variable I_1 can exist. Because of the supervenience relation between *M1* and *P1*, any variable I_1 that causes a change in *M1* also causes a change in *P1* (S2) and this variable *P1* is on a causal path to *P2* that does not go through *M1* (S3). In other words, every time we perform an intervention on a subjects' mental state, by manipulating some variable in separate experimental conditions, we also intervene on their brain state. This is not because the change in mental state causes the change in brain state (recall that a supervenience relation is not a causal relation; S1), but because the intervention changes both the mental state and the brain state (S2). In other words, we cannot control the effect of *P1* on *P2*. It follows that we cannot draw any conclusions about the causal effect of the intervention on the mental state. Furthermore, because the relation between *M1* and *P1* is not a causal relation, we also cannot say that *M1* is a contributing cause to *P2*. In the context of an experiment, we would say that *P1* is a confounding variable for which we cannot control, prohibiting any statement to be made about the causal effect of the independent variable on the dependent variable.

Woodward's Response

In reply to Baumgartner's argument, Woodward (6) proposes that when assessing causation in a variable set that includes supervenience relations between variables, it is not necessary to control for or hold fixed the supervenience base. Thus, it is not necessary to control for *P1* when assessing the relation between *M1* and *P2*. According to Woodward, this is because the interventionist account of causation as defined by (M) and (IV) is intended to apply to systems of causal relations in which no non-causal relations (such as supervenience relations) exist. It is not at all clear whether it is applicable to a system in which non-causal relations are present.

Woodward illustrates this by giving an example of a variable set in which non-causal relations are present that are

not supervenience relations (6). His example goes along the following lines. Suppose that getting a headache (*H*) is causally influenced by the amount of alcohol consumption (*AC*), which increases the probability of getting a headache, and the amount of non-alcoholic liquid consumption (*NC*), which decreases the probability of getting a headache. We also have a variable representing the total liquid consumption (*TC*), which is the sum of *AC* and *NC*. Assume that we also think of *TC* as causally influencing *H*. We can put all these variables together to get the schematic representation in **Figure 3**.

Suppose now that we want to investigate if *AC* is causally relevant for *H*. According to Baumgartner's reading of (IV) and (M), this would mean that it has to be possible to change (intervene on) *AC* without changing any other variable in **Figure 4** that is on a directed path to *H* that does not go through *AC*. We can see that this is not possible because *TC* is defined such that if *AC* changes, *TC* also changes. It seems strange to take this finding as evidence for there not being a causal relation between *AC* and *H*. Therefore, Woodward (6) concludes, the interventionist definition as put forward in (M) and (IV) is intended to only apply to systems of causal relations in which no non-causal relations exist. In systems with non-causal relations, one needs to hold fixed only the *appropriate* variables. In the variable set described in **Figure 4**, this means that if one wants to investigate the effect of *AC* on *H*, *NC* needs to be fixed, but *TC* does not.

Similarly, Woodward (6) argues, when one wants to investigate the causal effects of supervening variables, their supervenience base does not have to be fixed. This means that *M1* can be causally relevant for *P2* or in other words, according to this interpretation of interventionist causation, investigating the relation between mental states and brain states, as done in traditional cognitive neuroscience studies, by manipulating *M1* and investigating its effect on *P2* can reveal a causal relation between *M1* and *P2*.

In conclusion, according to Baumgartner (9), one cannot draw any conclusions about causal relations between mental states and brain states from traditional cognitive neuroscience studies. However, according to Woodward (6), this is perfectly valid. In the next section, we will discuss both these positions in light of brain stimulation studies in which the brain states are manipulated to investigate the effects on mental states.

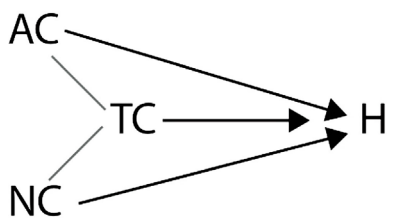


FIGURE 3 | Schematic representation of the relation between AC, WC, TC, and H.

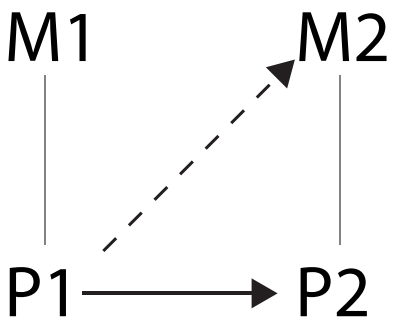


FIGURE 4 | Schematic representation of the relation investigated in brain stimulation studies as indicated by the dashed arrow.

CAUSATION IN BRAIN STIMULATION STUDIES

Since the introduction of brain stimulation techniques such as TMS and DBS, it has become possible for scientists to directly manipulate (intervene on) the electrical activity in the brain. Many neuroscientists have been using these techniques to draw conclusions about the causal relations between brain states and mental states. The following are quotes from TMS and DBS studies published in high-impact journals:

> "Making the causal link: frontal cortex activity and repetition priming" (10).
> "Causal implication by rhythmic TMS of alpha frequency in feature-based vs. global attention" (11).
> "DBS of the subthalamic nucleus markedly improves the motor symptom's of Parkinson's disease, but causes cognitive side effects such as impulsivity" (12).
> "Stimulation of a restricted site in the upper midbrain can cause major acute depression" (13).

Are these claims justified? In the present section, we will explore this question in light of Baumgartner's and Woodward's arguments. Before we continue, we should mention that a large part of the brain stimulation studies only focuses on the effects of changes in brain states on other brain states [e.g., Ref. (14–16)]. This is the relation between *P1* and *P2*. As described in (S3), we assume that there exists a causal relation between these variables. Therefore, in these types of brain stimulation studies, it is perfectly justified to talk about causal effects.

In the brain stimulation studies in which the relation between brain states and mental states is investigated, this seems more complicated. In these studies, the relation between *P1* and *M2* as depicted in **Figure 4** is investigated. *P1* is manipulated by stimulating a certain brain area using TMS or DBS in one condition and not stimulating it in another condition, while measuring some mental variable *M2* in both conditions. The researcher tries to make sure that the two conditions only differ on *P* and not on other variables, for example by applying sham stimulation in the control condition. If then a (significant) difference in *M2* between the two conditions is found, the researcher concludes that there was an effect of the change in brain activity on the mental state. However, is he or she justified in saying that *P1* has *caused M2*?

Baumgartner's Approach

To determine whether Baumgartner's argument applies to this experimental set-up, the conditional (BM) has to be redefined. If we switch the relevant terms, we get the following definition:

> (BP) If *P1* is causally relevant to *M2* with respect to the variable set $\mathbf{V} = \{M1, M2, P1, P2\}$, then there possibly exists a variable I_1 that causes a change in the value (or the probability distribution) of *P1* and is statistically independent of any variable *Z* that causes *M2* and that is on a directed path that does not go through *P1*.

Interestingly, the causal exclusion argument used by Baumgartner (9) to conclude that *M1* is not causally relevant to *P2* does not work in this case. This is because there is no variable in **V** that causes *M2* but does not go through *P1*. It is true that, because of the supervenience relation (S2), any intervention on *P1* also changes *M1*. However, there is no causal relation between *M1* and *M2* that does not go through *P1*. Furthermore, an intervention on *P1* also changes *P2,* through the causal relation mentioned in (S3), but according to (S1), *P*2 does not cause *M2*. So it seems that an intervention on *P1* is possible without intervening on another variable that causes *M*2. This suggests that we actually can make causal claims on the basis of brain stimulation studies, even if we assume a supervenience relation and accept Woodward's original intervention criteria. Let us now see whether Woodward's approach leads to a similar conclusion.

Woodward's Approach

According to Woodward, if one wants to investigate whether *P1* is causally relevant to *M2*, one needs to perform an intervention to change the value of *P*, while holding fixed all *appropriate* other variables. When investigating the relation between *P1* and *M2*, the supervenience base *P2* is not one of these appropriate variables, so even if *P2* were on a directed path to *M2* that does not include *P1*, *P2* does not have to be fixed because it is the supervenience base of *M*2.

The other possible candidate for a variable that is on a directed path to *M2* that does not include *P1*, is *M1*. Now this seems to pose a problem. According to Woodward's adjusted interpretation of interventionist causation, we can argue that *M1* causes *M2*, because *P1* and *P2* do not have to be fixed. This seems to imply that *M1* is an alternative cause for *M2* making it impossible to conclude that *P1* has caused *M2*. However, it seems that in his adaptation, Woodward (6) also argues that supervening variables do not have to stay fixed:

> (IV*) An intervention *I* on *X* with respect to *Y* will (a) fix the value of *SB(X)* in a way that respects the supervenience relationship between *X* and *SB(X)*, and (b) the requirements in the definition (IV) are understood as applying only to those variables that are causally related to *X* and *Y* or are correlated with them *but not to those variables that are related to X and Y as a result of supervenience relations* [(6), pp. 32].

This means that *M1* does not have to be fixed in order to draw a causal conclusion about the relation between *P1* and *M2* by intervening on *P1*.

Thus, according to Woodward's adjusted interpretation of interventionist causation, brain stimulation studies in which appropriate controls are applied, such as randomization of groups and application of sham stimulation in the control group, are suitable to base conclusions about the causal effect of brain states on mental states on.

In conclusion, if one follows Woodward, we can make claims about causal relations between brain states and mental states from the results of both traditional cognitive neuroscience and brain stimulation studies. However, for this to work, we do have to adjust the original interventionist criteria (5) and accept a non-causal supervenience relation. According to Baumgartner, by contrast, we cannot make claims about causation from the results of traditional cognitive neuroscience studies. However, even if we do not adjust the original criteria, we can still draw conclusions about causation from brain stimulation studies.

ARTICULATING THE MIND–BRAIN RELATION

The conclusions drawn in the previous sections rely heavily on the assumption of a supervenience relation between brain states and mental states. We believe that most neuroscientists would agree with this assumption. Quoting one of the key textbooks in cognitive neuroscience programs: "Cognitive neuroscience is an academic field concerned with the scientific study of biological substrates underlying cognition, with a specific focus on the neural substrates of mental processes" [(17), p. 12]. This definition suggests that what lies at the heart of cognitive neuroscience is a dependency relation between mental states ("cognition") and brain states (the "neural substrate").

Supervenience is not a "deep" explanatory relation; however, it only indicates the presence of a dependence relation without telling us what it is (18). A common way to further explain this dependency is by appealing to the notion of *emergence*. Central to emergentism is the idea that supervenient properties are "novel" properties over and above the properties upon, which they supervene. In the context of the mental causation debate, emergentism can be understood as the more specific claim that mental states are the emergent properties of a complex physical system, which have their own causal power and cannot be reduced to the basic physical properties of this system.[2] It is also possible to explain the dependence relation between mental and physical properties in terms of *reduction*. In contrast to emergentism, reductionism claims that supervenient properties are reducible to their base properties, and hence that mental properties are reducible to physical properties.

Thus far we have investigated whether interventionism allows us to make causal claims about the relation between mental states

[2] See, e.g., Ref. (19, 20) for specific accounts of supervenient emergentism in (cognitive) neuroscience.

and brain states, given a notion of supervenience that in principle allows mental states to be causally efficacious *qua* mental. In other words, we have assumed a dependence relation between mental states and brain states that is non-reductive, in principle compatible with the notion of emergence, and less strong than an identity relation. However, some people might not agree with this characterization of the relation between mind and brain. What happens to the conclusions drawn in this paper when one wishes to assume a reductive relation between mental states and brain states instead? In what follows we will briefly explore this option and also consider the possibility of a causal relation.

Type Identity and Functional Reduction

A radical reductive explanation of the supervenience relation is offered by the identity theory. This theory holds that mental states *are* brain states. The strongest version of the identity theory, the so-called "type-identity" theory, is reductionist in the sense that it states that specific types of mental states can be reduced to specific types of brain states (21). This theory would therefore claim that $M1 = P1$ and $M2 = P2$. Now it becomes almost trivial to show that an intervention on *P1* can show a causal effect on *M2*: since we know that an intervention on *P1* will cause a change in *P2* (S3) and since *M2* is now the same as *P2,* we can conclude that *P1* causes *M2*. Thus, assuming a type-identity relation between brain states and mental states still allows us to draw conclusions about the causal effects of brain states on mental states from brain stimulation studies. Similarly, assuming a type-identity relation also makes it possible to draw conclusions about causal relations from traditional cognitive neuroscience studies: since we know an intervention on *P1* causes a change in *P2* and $M1 = P1$, we can conclude that *M1* causes *P2*.

The identity theory faces two important problems. First of all, it does not really provide us with an explanation of *why* mental states are identical with brain states. Take the claim that water is H_2O. In this case, we can explain the properties of water in terms of the molecules that constitute it (two hydrogen atoms and a single oxygen atom) and the way they are interrelated. Stating that mental states are identical with brain states does not provide us with such an explanation. Second, there is the problem of multiple realizability (22, 23). If (at least some) mental states can be realized by different brain states, which seems plausible given what we know about the plasticity of the human brain, then they cannot be identical with specific brain states.

An alternative model of reduction, *functional reduction*, has been proposed by Kim (18, 24). According to this model, mental states can be reduced in the following way:

> *Stage 1.* Define *M* in terms of its "causal role," i.e., in terms of the causal task *C* it performs.
>
> *Stage 2.* Identify the "realizers" of *M*, i.e., the actual mechanisms that perform causal task *C*.
>
> *Stage 3.* Develop an explanatory theory that explains how the realizers of *M* perform causal task *C*.

The causal claims about brain stimulation studies and traditional cognitive neuroscience studies that can be made on the basis of functional reductionism are similar to those that can be made on the basis of the identity theory. Furthermore, functional reductionism does provide an explanation of how mental states are realized by brain states, and it is entirely consistent with the phenomenon of multiple realizability (in the sense that Stage 2 anticipates the existence of multiple lower-level realizers).

However, functional reductionism, like the identity theory, comes at a high price: it grants mental states causal power, but only in virtue of their being physical states. And this might be a hard pill to swallow, since many people believe that mental states do have causal power of their own, *qua* mental. It is precisely this intuition that is behind the debate between Baumgartner and Woodward in the first place.

A Causal Relation

The second alternative that we will consider is one that postulates a causal relation between brain states and mental states, in the sense that *P2* causes *M2*. Although most philosophers reject this possibility [with the notable exception of (25)], it might strike cognitive neuroscientists as a plausible option.

What can we conclude from brain stimulation studies if we assume that the relation between *P2* to *M2* is a causal relation (one that has been established by means of an appropriate intervention)? In particular, can we still make causal claims about the relation between *P1* and *M2*? At first glance, the problem seems to be that *P2* is now on a directed path to *M2* that does not include *P1*. However, it is precisely this relation that allows us to conclude that *P1* is a *contributing* cause according to the second part of (M):

> A necessary and sufficient condition for *X* to be a (type-level) *contributing cause* of *Y* with respect to variable set **V** is that (i) there be a directed path from *X* to *Y* such that each link in this path is a direct causal relationship… and that (ii) there be some intervention on *X* that will change *Y* when all other variables in **V** that are not on this path are held fixed [(5), pp. 59].

Note that this was not possible when we assumed a supervenience relation between *P2* and *M2* because in that case not every link on the path from *P1* to *M2* was a direct causal relation.

Unfortunately, this does not work when we want to make causal inferences from traditional cognitive neuroscience studies. *M1* cannot have a causal effect on *P2*, because it is impossible to intervene on *M1* without violating the second requirement of (IV):

> 2 *I* acts as a switch for all the other variables that cause *X*. That is, certain values of *I* are such that when *I* attains those values, *X* ceases to depend on the values of other variables that cause *X* and instead depends only on the value taken by *I*.

Criterion 2 is violated because of the assumption that *M1* is caused by *P1*. What this seems to show is that the assumption of a causal relation between brain states and mental states ultimately leads to *epiphenomenalism*, i.e., the thesis that mental states can be causally influenced by physical states, but have no causal efficacy themselves. It is probably safe to assume that most people will not really consider this an improvement over the minimal notion of mental causation provided by type identity and functional reduction.

CONCLUSION

The aim of this paper was to investigate whether we can draw conclusions about causal relations between brain states and mental states from traditional cognitive neuroscience studies and brain stimulation studies, given an interventionist account of causation. We have argued that, if one follows Woodward in embracing the notion of supervenience and revising the criteria for what counts as an intervention, both types of studies can be used to establish causal claims. If, by contrast, one follows Baumgartner and his causal exclusion argument, traditional cognitive neuroscience studies cannot be used to establish causal claims but brain stimulation studies can.

Brain stimulation is being used more and more as a form of treatment for psychiatric disorders. We have shown that from an interventionist point of view, it is valid to say that these brain stimulation treatments *cause* changes in mental states. Now this is not necessarily an argument in favor of these treatments. However, if Baumgartner is right, then it seems reasonable to conclude that brain stimulation treatments will become increasingly attractive. Unlike traditional cognitive neuroscience studies, they actually have the potential to elucidate the causal structure of certain psychiatric disorders (i.e., the underlying causal relations between mental states and brain states). It is a safe bet that this will appeal to many psychiatrists.

At the same time, this conclusion is based on a "conservative" interpretation of interventionism, and a rejection of non-causal metaphysical relations between mental states and brain states such as supervenience. In this respect, Woodward's position is much more "liberal," insofar as it proposes adjusted intervention criteria and allows for the inclusion of non-causal relations between mental states and brain states. One advantage of Woodward's position is that it allows psychiatrists to draw conclusions about the causal structure of mental disorders from traditional cognitive neuroscience (and not just brain stimulation studies). Another and perhaps even more important advantage is that it legitimates the claim that cognitive and behavioral therapy, aiming at influencing the mental state of a patient, can *cause* changes in the patient's brain state.

A note of caution is required when applying our conclusions to psychiatric practice. When defining mental states and brain states as separate targets of intervention, we assume an ideal situation in which such a separation can be easily obtained, and Woodward's intervention criteria are met. In practice, however, such a situation might be difficult to achieve. For example, in their paper on degeneracy, Price and Friston (26) have argued that different neural configurations can lead to similar mental states. This means that a disruption of one of these configurations by brain stimulation might not necessarily lead to a change in mental state. The experimenter who uses interventionism in the context of a single study would then be forced to conclude that there is no causal relation between the brain state and mental state in question. Now this scenario could be avoided by making sure that conclusions about causal relations between mental states and brain states are supported by multiple studies (controlling for both inter- and intra-individual variation). However, there might be a larger worry here, not just about the fact that the application of interventionism to single studies in practice might sometimes result in misguided causal claims, but also about the *very possibility* of applying interventionism to cognitive neuroscience studies.

For example, one might argue that various factors such as degeneracy, redundancy, path-dependency, non-linearity and complex feedback loops make it (theoretically) impossible to establish linear causal chains.[3] In the light of this, several theorists have proposed a concept of "circular causation" (27–29). Circular causation, which is taken to be typical of a self-organizing system, is realized by the cooperation of the individual parts of the system, yet it also governs or constrains the behavior of these individual parts. A good illustration of circular causation is given by McGilchrist (30) in his account of the brain as a complex system: "Events anywhere in the brain are connected to, and potentially have consequences for, other regions, which may respond to, propagate, enhance or develop that initial event, or alternatively redress it in some way, inhibit it, or strive to re-establish equilibrium. There are no bits, only networks, an almost infinite array of pathways" (2010, p. 34).

Circular causation is attractive, but also slightly misleading – at least when it is articulated in opposition to linear causation. As Von Bertalanffy (31, 32) already pointed out, to make sense of circular causation we still require a notion of linear and "unidirectional" causation. The kind of feedback regulation that is implied by circular causation is obviously not unidirectional in *spatial* terms: it moves back and forth or circles around the various components of a system (33). However, despite circling in space, feedback still proceeds forward in *linear time*, one component being separated from the next in time. Circular causation, thus understood, is compatible with the assumption of a supervenience relation between mental states and brain states. The question is whether it is also compatible with interventionism. Let us say we propose a modified version of **Figure 1** that involves feedback loops, for example one in which P1 causes P2 which in turns causes *P1* (**Figure 5A**). Now, at first glance, such a feedback loop seems to violate the second requirement of (IV), in the sense that one might think that *P1* not only depends on the value taken by *I* but also on the value of *P2*. However, the problem is that such a depiction of circular causation fails to take into account the fact

[3]We thank an anonymous reviewer for bringing this to our attention.

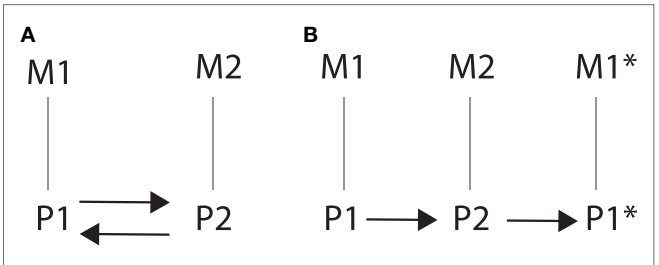


FIGURE 5 | **(A)** Incorrect depiction of circular causation that fails to take into account the temporal relation between P1 and P2, and P2 and P1. **(B)** Correct depiction of circular causation that takes into account the spatial as well as the temporal relation between P1 and P2, and P2 and P1*, and illustrates the linear nature of circular causation.

that the relations between *P1* and *P2*, and *P2* and *P1* are temporal relations between different *events,* and therefore they cannot be circular. That is, the *P1* that causes *P2* is different (not spatially, but temporally) from the *P1* that is caused by *P2* as the result of the feedback loop. The correct (linear) way to represent circular causation is shown in **Figure 5B**: *P1* causes *P2*, and *P2* causes *P1** (which is temporally different, but spatially identical with *P1*). And this seems to be compatible with interventionism, to the extent that an intervention on *P1* does not violate the second requirement of (IV).

It is important to note, at this point, that interventionism as such is relatively free of metaphysical commitments, in the sense that it does not make claims about how exactly one should spell out the relation between mental states and brain states. It only tells us what needs to be in place and which conditions need to be met for a given relation between variables to be described as "causal." Furthermore, as we have shown in Section "The Interventionist Account of Causation," one of the main attractions of interventionism is that it seems to correspond to how causal relations are investigated in scientific practice. Therefore, even if interventionism turns out to be incompatible with certain assumptions about brain functioning, such as circular causation and feedback loops, then this indicates a larger problem with mainstream scientific method and textbook accounts of experimental research. Obviously, these are issues that deserve critical attention. For the purpose of this paper, however, we have taken the mainstream scientific method as our starting point.

In the end, how psychiatrists approach these issues will probably depend on their intuitions about interventionist causation and the relation between mind and brain. However, one thing is certain. What they conclude will have important implications for the way they communicate the effect of different treatments to their patients.

AUTHOR CONTRIBUTIONS

ND wrote the core sections of this paper ("The Interventionist Account of Causation," "Mental States and Brain States: a Supervenience Relation," "Causation in Traditional Cognitive Neuroscience Studies," and "Causation in Brain Stimulation Studies") and the main argument; LB wrote the other Sections "Introduction," "Articulating the Mind–Brain Relation," and "Conclusion," revised/rewrote Sections "The Interventionist Account of Causation," "Mental States and Brain States: a Supervenience Relation," "Causation in Traditional Cognitive Neuroscience Studies," and "Causation in Brain Stimulation Studies and supervised the project."

FUNDING

LB's research was supported by a grant from the Templeton World Charity Foundation. The opinions expressed in this publication are his own and do not necessarily reflect the views of Templeton World Charity Foundation.

REFERENCES

1. Sack AT. Transcranial magnetic stimulation, causal structure-function mapping and networks of functional relevance. *Curr Opin Neurobiol* (2006) 16(5):593–9. doi:10.1016/j.conb.2006.06.016
2. Cleary DR, Ozpinar A, Raslan AM, Ko AL. Deep brain stimulation for psychiatric disorders: where we are now. *Neurosurg Focus* (2015) 38(6):E2. doi:10.3171/2015.3.focus1546
3. Holtzheimer PE, Mayberg HS. Deep brain stimulation for psychiatric disorders. *Annu Rev Neurosci* (2011) 34:289–307. doi:10.1146/annurev-neuro-061010-113638
4. Mayberg HS, Lozano AM, Voon V, McNeely HE, Seminowicz D, Hamani C, et al. Deep brain stimulation for treatment-resistant depression. *Neuron* (2005) 45(5):651–60. doi:10.1016/j.neuron.2005.02.014
5. Woodward J. *Making Things Happen: A Theory of Causal Explanation*. (Vol. 14). Oxford: Oxford University Press (2007).
6. Woodward J. Interventionism and causal exclusion. *Philos Penomenol Res* (2014). doi:10.1111/phpr.12095
7. Convenor C, Field A, Drury J. *Research Mewwthods in Psychology*. (Vol. 26). (2006). Available from: www.Sussex.ac.uk.
8. Shaughnessy JJ, Zechmeister EB, Zechmeister JS. *Research Methods in Psychology*. New York: McGraw-Hill (2012).
9. Baumgartner M. Interventionist causal exclusion and non-reductive physicalism. *Int Stud Philos Sci* (2009) 23(2):161–78. doi:10.1080/02698590903006909
10. Martin A, Gotts SJ. Making the causal link: frontal cortex activity and repetition priming. *Nat Neurosci* (2005) 8(9):1134–5. doi:10.1038/nn0905-1134
11. Romei V, Thut G, Mok RM, Schyns PG, Driver J. Causal implication by rhythmic transcranial magnetic stimulation of alpha frequency in feature-based local vs. global attention. *Eur J Neurosci* (2012) 35:968–74. doi:10.1111/j.1460-9568.2012.08020.x
12. Frank MJ, Samantha J, Moustafa AA, Sherman SJ. Hold your horses: impulsivity, and medication in parkinsonism. *Science* (2007) 318(5854):1309–12. doi:10.1126/science.1146157
13. Bejjani B, Damier P, Arnul I, Thivard L, Bonnet A, Dormont D, et al. Transient acute depression induced by high-freuquency deep-brain stimulation. *N Engl J Med* (1999) 340:1476–80. doi:10.1056/NEJM199905133401905
14. Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, et al. Brain areas involved in perception of biological motion. *J Cogn Neurosci* (2000) 12(5):711–20. doi:10.1162/089892900562417
15. Thut G, Miniussi C. New insights into rhythmic brain activity from TMS-EEG studies. *Trends Cogn Sci* (2009) 13(4):182–9. doi:10.1016/j.tics.2009.01.004
16. Zatorre RJ, Chen JL, Penhune VB. When the brain plays music: auditory-motor interactions in music perception and production. *Nat Rev Neurosci* (2007) 8(7):547–58. doi:10.1038/nrn2152
17. Gazzaniga MS, Irvy RB, Mangun GR. *Cognitive Neuroscience: The Biology of the Mind*. New York: W.W. Norton (2002).
18. Kim J. Supervenience, emergence, realization, reduction. In: Loux M, Zimmerman D, editors. *The Oxford Handbook of Metaphysics*. Oxford: Oxford University Press (2005). p. 556–84.
19. Haken H. *Brain Dynamics*. Berlin: Springer (2002).
20. Atmanspacher H. Identifying mental states from neural states under mental constraints. *Interface Focus* (2012) 2:74–81. doi:10.1098/rsfs.2011.0058
21. Smart J. *Philosophy and Scientific Realism*. London: Routledge (2014).
22. Putnam H. Psychological predicates. In: Capitan W, Merrill D, editors. *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press (1967). p. 37–48.
23. Fodor J. Special sciences: or the disunity of science as a working hypothesis. *Synthese* (1974) 28:97–115. doi:10.1007/BF00485230
24. Kim J. *Mind in a Physical World*. Cambridge, MA: MIT Press (1998).
25. Searle JR. Reductionism and the irreducibility of consciousness. In: Flanagan O, Block N, Guzeldere G, editors. *The Nature of Consciousness*. MIT Press (1997).
26. Price CJ, Friston KJ. Degeneracy and cognitive anatomy. *Trends Cogn Sci* (2002) 6:416–21. doi:10.1016/s1364-6613(02)01976-9
27. Varela F. *Principles of Biological Autonomy*. New York, NY, North Holland: Elsevier (1979).
28. Kelso JAS. *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge: MIT Press (1995).

29. Freeman WF. Consciousness, intentionality, and causality. *J Conscious Stud* (1999) 6:143–72.
30. McGilchrist I. *The Master and His Emissary: The Divided Brain and the Making of the Western World*. New Haven: Yale University Press (2009).
31. Von Bertalanffy L. *General Systems Theory*. New York: George Braziller (1968).
32. Von Bertalanffy L. General theory of systems: application to psychology. In: Kristeva J, Rey-Debove J, Umiker DJ, editors. *Essays in Semiotics*. The Hague: Mouton (1971).
33. Slife BD. *Time and Psychological Explanation*. New York, NY: State University of New York Press (1993).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 11 March 2016
doi: 10.3389/fpsyt.2016.00032

Explanatory Pluralism and the (Dis)Unity of Science: The Argument from Incompatible Counterfactual Consequences

*Victor Gijsbers**

Institute for Philosophy, Universiteit Leiden, Leiden, Netherlands

What is the relationship between different sciences or research approaches that deal with the same phenomena, for instance, with the phenomena of the human mind? Answers to this question range from a monist perspective according to which one of these approaches is privileged over the others, through an integrationist perspective according to which they must strive to form a unity greater than the sum of its parts, to an isolationist perspective according to which each of them has its own autonomous sphere of validity. In order to assess these perspectives in this article, I discuss the debates about the unity of science and about explanatory pluralism. The most pressing issue turns out to be the choice between the integrative and the isolationist perspective: the question is whether the integrative tendencies in science should be fully indulged in or whether they should be held in check by acknowledging that a certain amount of isolation is necessary. I argue that the issue can be further distilled into the question of whether two true explanations of the same fact can ever fail to be combinable into one single explanation. I show that this can indeed be the case, namely, when the explanations have incompatible counterfactual consequences, something that is often the case when we try to combine explanations from different sciences or research approaches. These approaches thus embody perspectives on the world that are to a certain extent autonomous. This leads to the conclusion that although interdisciplinarity may have many advantages, we should not take the project of integration too far. At the end of the day, the different research approaches with their different perspectives and insights must remain precisely that: different and somewhat disunified.

Keywords: explanatory pluralism, unity of science, disunity of science, explanation, counterfactual incompatibility, counterfactuals

OPEN ACCESS

Edited by:
Leon De Bruin, VU University Amsterdam, Netherlands

Reviewed by:
Markus Ilkka Eronen, KU Leuven, Belgium
Jeroen Van Bouwel, Ghent University, Belgium

***Correspondence:**
Victor Gijsbers
V.Gijsbers@hum.leidenuniv.nl

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry

Received: *25 October 2015*
Accepted: *25 February 2016*
Published: *11 March 2016*

Citation:
Gijsbers V (2016) Explanatory Pluralism and the (Dis)Unity of Science: The Argument from Incompatible Counterfactual Consequences. Front. Psychiatry 7:32. doi: 10.3389/fpsyt.2016.00032

INTRODUCTION

What is the relationship between the different sciences or, to use a more fine-grained term, research approaches that deal with the human mind? Faced with a variety of explanations for a psychiatric illness – for example, genetic, neurological, cognitive, psychoanalytic, and sociological explanations – a scientist could take one of three broad views. First, the view that one explanation will trump all the others, making it the only one that is needed. Second, the view that these explanations can be

combined into one unified explanation that is superior to each of the individual ones. Third, the view that these explanations all add to our understanding, but cannot be combined into a single integrated account. In other words, a scientist could take a *monist*, an *integrationist*, or an *isolationist* view. Of course, combinations are also possible: one might believe that genetic and neurological explanations can be integrated, that sociological explanations are useful but will remain isolated, that explanations from cognitive psychology will be trumped by neurological ones, and that psychoanalytic explanations are simply wrong; or any other combination of options.

The distinction between these views has important practical consequences both for research and for therapy. If a monist perspective were correct, scientists and practitioners alike should learn to focus on the specific approach that provides the best explanations (which monists of today are likely to identify with neuroscience and low-level biological approaches to the human brain). But if an integrationist perspective were correct, researchers should focus on doing interdisciplinary research, and practitioners should make sure that they learn to see the interconnections between different clinical approaches and find ways to combine them. Then again, if an isolationist perspective were correct, the most fruitful approach would be one of disciplinary specialization and parallel but isolated lines of treatment.

The general heading under which these issues used to be discussed was that of the *unity of science*, but as I will make clear in Section "From Unity of Science to Explanatory Pluralism," philosophers now generally talk about the issue of *explanatory pluralism*. Those who defend explanatory pluralism – which includes almost everyone engaged in the current discussion – generally reject the monist and isolationist positions in favor of some kind of integrationism, some version of the idea that the different sciences have to work together in order to achieve results that they could not achieve separately. The arguments for this claim come in two flavors: first, arguments to the effect that interdisciplinary research is methodologically superior to monodisciplinary research, and, second, case studies that prove that scientists are actually pursuing such research and achieving their aims through it.

These methodological and empirical approaches are of course very valuable. But in the current paper, I would like to ask a more fundamental question about scientific explanation, namely, whether there is anything in the structure of explanations itself that can form a barrier to their complete integration. If two different research approaches come up with explanations of the same phenomenon, and if these are both true and not logically contradictory, is it then always possible to put them together into a single integrated perspective on the phenomenon? Or is it sometimes the case that we have no choice but to be isolationists, because the explanations themselves just do not fit together? Can true explanations be incompatible? These questions are pertinent. There may be, in our current scientific practice, a presumption in favor of interdisciplinarity and the integration of explanations; but it is far from clear that complete integration is the correct ideal to pursue. Perhaps there is a sense in which, say, a neurological and a sociological explanation of a patient's symptoms are just "too different" to be forged into a single, more complete explanation.

My task in Section "Combining Explanations" will be to analyze the conditions under which two explanations of the same phenomenon can fail to be combinable into a single explanation; in other words, I want to find out what it would mean for two explanations to be "too different." To this end, I will develop the notion of *counterfactual incompatibility*: the idea that two statements, even though they are logically consistent, can nevertheless imply different things about what would have happened under hypothetical circumstances. I then argue that explanations that are counterfactually incompatible cannot be combined into a single explanation – and far from that being a merely academic possibility, this does in fact regularly happen when we take explanations from different research approaches.

We thus find, from studying the structure of explanations themselves, that there is something in the sciences that resist integration; that, however, much we love interdisciplinarity, there is an extent to which we must remain isolationists; that different research approaches yield perspectives on the world which cannot always be fully integrated. Interdisciplinarity and the search for connections, while commendable, should be held in check by a healthy appreciation of the autonomy of each of the individual approaches scientists are using.

FROM UNITY OF SCIENCE TO EXPLANATORY PLURALISM

The 1930s saw a rising interest in the idea of the unity of science, which is perhaps nowhere more visible than in the activities of one of the fathers of logical empiricism, Otto Neurath. Neurath founded the Unity of Science Institute in 1936; organized a series of conferences between 1935 and 1941 called the International Congresses for the Unity of Science; and started the International Encyclopedia of Unified Science [see Cat (1) for historical details]. Many of the most important philosophers of the era were implicated in one or more of these enterprises, among them Philipp Frank, Charles Morris, Rudolf Carnap, Bertrand Russell, Ernest Nagel, and John Dewey.

Neurath's own overarching concern with the unity of science movement was to create an environment within which the different sciences could interact and learn from each other. As Pombo et al. (2) put it:

> Neurath's own encyclopedic conception of the unity of science is built on the notion of cooperative action in the scientific community and the accumulation of available results. [...] at the heart of the project is the goal of providing a universal medium for communicating across disciplines and languages (p. 4)

But although Neurath's own view was characterized by symmetric and non-reductionist ideas about communication, cooperation, interdisciplinarity, and interaction between different disciplines (idem, p. 6), the idea of the unity of science was soon interpreted in a much more reductive way. Thus, Oppenheim & Putnam (3) suggests that the unity of science would be achieved when all the terms and all the laws of all the sciences have been

reduced to the terms and laws of a single scientific discipline. For Oppenheim and Putnam, the unity of science in this sense is an "over-arching metascientific hypothesis" (p. 6) which, even if it cannot be conclusively shown to be true, is nevertheless credible (p. 8).

It is against this background that we have to understand the position taken up in Fodor's classic anti-reductionist paper "Special Sciences, or: the Disunity of Science as a Working Hypothesis" (4), namely, the position that it is in all probability useless to search for lawful coextension of predicates from sciences at different levels. If unity of science is understood in the reductionist way that Oppenheim and Putnam understand it, then one is indeed tempted to emphasize, with Fodor, the *disunity* of science rather than its unity. Few scientists are interested in strong reductionist projects, and thus it might seem that they have no reason to seek for a unity of science.

In the philosophy of science, this attitude has been forcefully defended by Dupré (5), Cartwright (6), and Teller (7). They use metaphysical, epistemological, and methodological arguments to argue for a "disunified" and "dappled" view of science in which no overarching, all-encompassing laws can be found, and no single discipline will emerge as foundational. Dupré, in fact, insists that science is not even unified by any common sociological, methodological, or processual element.

We have no quarrel with this general view, but something seems to be missing from it. We are still, and perhaps more than ever, interested in communication, cooperation, interdisciplinarity, and interaction between the sciences; and we share with Oppenheim and Putnam, if not their views about reduction, at least their general aim of "counterbalancing specialization by promoting the integration of scientific knowledge." Insisting, as Dupré, Cartwright, and Teller do, that science cannot be unified, helps combat reductionist ideals, but does little to shed light on why integration is still seen as a worthy goal.

How can we understand the unifying tendency in science without returning to reductionism? One option is to defend non-reductive unity at a metaphysical level: examples of this are non-reductive physicalism and the idea of the "primacy of physics" defended by Ladyman and Ross (8). Such metaphysical discussions will be avoided in the current paper, in order to focus on the methodology and the products of science. Among methodologically inclined philosopher of science, there seems to be an emerging tendency to revive a Neurathian use of the term "unity of science," as some of the authors in Symons et al. (9) do. But more influential, especially among scientists themselves, has been the adoption of a new term of art for what is at bottom the same idea of unifying different research approaches: explanatory pluralism. It is this term and the debate surrounding it that we will focus on.

The term "explanatory pluralism" is not without problems, one of which is that different authors use it in sometimes quite different ways. But we can glean the core idea from some representative citations, all of them from papers which set out to defend a form of explanatory pluralism within the psychological sciences:

> Explanatory pluralism holds that simultaneously pursuing research at multiple analytical levels in science tends to aid progress at each of those levels [(10), p. 738]
>
> Explanatory pluralism hypothesizes multiple mutually informative perspectives with which to approach natural phenomena [(11), p. 436]
>
> On this view, different sciences have a degree of autonomy (they are not to be eliminated), but also interact in an effort to understand physical reality at different scales (they are not fully autonomous silos). [...] different sciences and theoretical approaches should maintain their emphasis on different proprietary scales but should also work to unify their work as much as possible, insofar as they often describe the same phenomena in different but compatible ways [(12), p. 3]

As we can see, there are two elements to the core idea: first, that science can be broken up into distinct enterprises, and, second, that it is scientifically fruitful to have interaction between these enterprises. The authors have different ideas about how to carve up science: in terms of "levels," or "perspectives," or "scales," or simply "sciences"; but in each case, we are presumably to identify them with well-known disciplines and subdisciplines, such as high-energy physics, cell biology, neuroscience, and cognitive psychology. We will ignore the question of how exactly to carve up science and assume that speaking of research approaches is clear enough.

More relevant to our current purposes is the second element of explanatory pluralism, namely, the idea that the sciences have to interact in order to achieve their full potential. As we can see, this claim is formulated in different ways by the different authors. McCauley and Bechtel frame it as a prediction about the rate of scientific progress, although one that is not, perhaps, especially clear, since it is not evident which contrast they are drawing. Kendler formulates explanatory pluralism as a methodological norm and contrasts it with reductionism. Abney et al. take explanatory pluralism to be an alternative not only to reductionism but also to an isolationist view of science that they attribute – perhaps inaccurately, since his article opposes reduction but not interaction in general – to Fodor (4). Given our interest in finding a middle ground between the unity and the disunity of science, this is an especially interesting version of the explanatory pluralism. But the formulation of Abney et al. remains vague: it exhorts us to unify the sciences "as much as possible," but it does not indicate how far that possibility extends.

Some of the most thoughtful analyses of explanatory pluralism are those of Marchionni (13), Mitchell (14), Campaner (15), and Van Bouwel (16). [Closely related, though couched in a different terminology and less focused on the technical details of explanation, is Brigandt (17) account of *explanatory integration* as an intermediate between reductionism and pluralism.] All these authors take the view that explanatory pluralism is primarily about what *explanations* the best science will end up with, and more precisely about the question whether explanations from different research approaches can all be integrated into a coherent whole.

Marchionni (13) makes a distinction between two ways in which explanations of the same phenomenon on a macro and micro level can complement each other: *weak complementarity*, which holds when the two explanations are both legitimate and autonomous, but cannot be combined; and *strong complementarity*, which holds when the two explanations can be integrated into a whole that provides a better explanation than the two explanations did separately. If weak complementarity holds, we have two research approaches that are essentially independent; this is a disunified or isolationist view of science. When strong complementarity holds, our best understanding of the world is generated when two or more research approaches interact: this is a unified or integrationist view. We thus arrive at a gliding scale ranging from the ultimate unity that is reduction/monism, to the ultimate disunity that is weak complementarity/isolationism, with strong complementarity/integrationism in between.

However, the idea of strong complementarity involves a certain instability. On the one hand, it poses different, distinct research approaches; and on the other hand, it tells us that the results of these approaches can be put together to form a single picture of the world, a picture that is more enlightening than any of the separate pictures. But if the sciences are to be integrated so tightly and do not have an autonomous domain of knowledge wherein they reign supreme, in what sense can they still be said to be distinct? Do they not reveal themselves as merely different parts of the same one-and-only scientific discipline?

This seems to be the background against which Campaner (15), in an attempt to explain how different kinds of psychiatric explanation can be combined, asks the following pertinent questions about explanatory pluralism:

> Is there any underlying idea that some sort of complete explanatory picture can be – sooner or later – elaborated, or is some more radical form of pluralism advanced here? Is pluralism suggested here as only the acknowledgement of the existence and toleration of a diversity of current explanatory theories, or also as the idea that distinctive views will persist as such in the long run? In other terms, is actual plurality treated in this context as provisional and resolvable, or is the idea that renouncing pluralism would lead to some loss of explanatory information? (pp. 98–99)

Unlike Marchionni, who comes out in favor of strong complementarity, Campaner believes that the different types of explanation in psychiatry will turn out to be impossible to integrate into a single type of explanation. She points at the very different aims and interests of different actors in the field of psychiatry, and she argues that there is little reason to suppose that the explanations constructed to advance those different interests will coincide, even in the long run. According to her, we must be "open to the possibility that, at least in principle, explanatory pluralism can be a permanent state" (idem, p. 101), where explanatory pluralism is here understood – justifiably, but somewhat confusingly when compared to Abney et al. – as the isolationist rather than the integrationist position.

Van Bouwel (16), in a commentary on Campaner and using and expanding the earlier classification of Mitchell (14), adds another level of sophistication to the analysis. Next to explanatory reductionism, Van Bouwel distinguishes no fewer than five different kinds of explanatory pluralism, ranging from the more monistic to the more pluralistic:

1. *Explanatory reductionism*: there is a single privileged research approaches, and ultimately the best understanding of the world will be achieved when all the explanations from other approaches are reduced to this privileged approach.
2. *Temporary pluralism*: it is methodologically advisable to promote a temporary plurality of competing theories, as a means of achieving, in the end, one single unified theory that gives the best explanations.
3. *Integrative pluralism*: satisfactory explanations can only be generated by integrating the findings of different research approaches. (This is equivalent to always embracing Marchionni's idea of strong complementarity.)
4. *Interactive pluralism*: research approaches often generate satisfactory explanations by themselves, but it is also often – though not invariably – the case that the integration of explanations from different sciences leads to a better explanation. (This position, which is Van Bouwel's preferred position, posits a mixture of Marchionni's two kinds of complementarity.)
5. *Isolationist pluralism*: different research approaches generate very different kinds of explanation, which are all valid but cannot be integrated. (This is equivalent to always embracing Marchionni's idea of weak complementarity.)
6. *Anything goes pluralism*: all theories and perspectives are equally valid, and the greatest understanding of the world is achieved by an unlimited proliferation of theories and perspectives.

Van Bouwel is undoubtedly right when he suggests that it would be tough to defend either the idea that isolation is always correct or the idea that integration is always correct. Interactive pluralism, which decides on a case-to-case basis whether integration will succeed or whether isolation is needed, seems to be the most rational position. But in its relaxed wait-and-see attitude, it misses out on something that is more adequately captured by the admonitions of McCauley, Bechtel, Kendler, and Abney et al. all of whom push toward integration. There is a methodological presumption in science in favor of integration: where we *can* integrate, one feels, we *should* integrate; after all, pushing toward integration has led to many great advances.[1] The scientist who insists on the splendid isolation of her discipline will come under immediate suspicion for being, perhaps, too conservative. "Interdisciplinarity" remains a word with which one can woo funding agencies. In other words, we love the unity of science, we are striving toward the unity of science, and if we fail to achieve

[1]See also Andler (18), pp. 140–141, for an appraisal of why we cannot ignore the unifying tendencies in science. Of course, there are critics of integration and interdisciplinarity too, but I venture – although I have no hard data to back this up – that most of these critics have doubts about the *possibility* of integration, rather than about the *desirability* of integration where this is possible.

it – that is, more specifically, if we fail to achieve an integrative pluralism where all the sciences work together to create one single coherent explanation of every phenomenon – than there must be some particular obstacle in the way of that integration. It is that obstacle that I wish to consider. Is the scientist who believes that her explanations stand alone and cannot fruitfully be combined with those of other sciences automatically an unintelligent conservative, or are there circumstances under which it is rational to embrace an isolationist pluralism? What, we may ask more specifically, are the circumstances under which two explanations can fail to be combinable into a single, more complete explanation?

Answering that question will be the burden of the Section "Combining Explanations" of this paper. But before I embark on that project, it will be useful to mention Van Bouwel's own approach to this question and distinguish my project from his. According to Van Bouwel et al. (19):

> [e]xplanatory pluralism consists in the claims that (i) the best form (and level) of explanation depends on the kind of question one is willing to answer by the explanation and (ii) that in order to answer all explanation-seeking questions in the best way possible we will need more than one form (and level) of explanation (p. 36)

The approach championed by these authors, which also influences Gervais' (20) account of inter-level explanations, starts not from a phenomenon, to then ask whether different research approaches should cooperate in giving a single explanation of that phenomenon, but starts from the idea that different epistemic interests lead to different explanatory questions that are best answered by explanations involving different forms and levels. This more pragmatic approach to explanation leads to a natural answer to the questions I posed above: yes, one can say, it is rational to believe that isolation is sometimes the best strategy, because under some circumstances narrow isolated explanations are more conducive to our specific epistemic goals than grand integrative stories. [In their 2011 article, Van Bouwel et al. (19) are actually concerned with showing that reductive explanations have a place in science next to high-level explanations, but I take it that they would also agree with the approach to isolation I just outlined.] Actual examples of science can then be used to prove that scientists indeed choose between integration and isolation based on pragmatic and contextual factors.

I have no quarrel with such an approach. Suppose, for a moment, that there is indeed a single best, completely integrated explanation of any phenomenon. Then, it is undoubtedly true – and I would expect even hard reductionists to agree – that there are strong pragmatic reasons against using this explanation to answer any and all questions about that phenomenon. A therapist interested in curing her patient's depression might not need to hear about the details of the patient's neurochemistry in order to prescribe the right cure, while the patient's company doctor might need to know nothing at all about the causes of the depression in order to decide whether or not to grant the patient extended sick leave. In practical contexts, the "best" explanation is often simple and idealized. And it would also be true, as the pragmatist might stress, that in practice we tend to lose important insights and information if we do not keep our practical goals in mind from the start, so that there is a more fundamental, if still practical, reason for pursuing isolated rather than integrated explanations.

So, even if it were true that there is a single best, completely integrated explanation of any phenomenon – where "best" is understood not in a pragmatic and contextual way, but in terms of an ideal state of understanding – there are still legitimate practical concerns about integration. But I want to know whether that supposition, which seems to underlie much of the theoretical defense of integration, is true. If it is, then the sciences are fundamentally one, at least as far as explanation is concerned; and we will reach the most perfect understanding of the world when we relentlessly pursue integration. If not, then the sciences are fundamentally a plurality; and we will lose some understanding if we push our quest for integration too far.

COMBINING EXPLANATIONS

Is there a single best, completely integrated explanation of any phenomenon? There are instances where one might doubt this for reasons having to do with what the explanations are *about*. For instance, one might doubt whether explanations involving the mind and explanations involving the body could ever be combined; or explanations involving facts and explanations involving values. These doubts are related to some of the thorniest metaphysical issues in all of philosophy. We will sidestep these issues – which we could not possibly do justice to here – and focus instead, not on what explanations are *about*, but on the general *form* or structure of explanations. What I want to know is what general feature of two explanations of the same phenomenon could stand in the way of their being combined into a single bigger explanation.

In order to simplify the discussion, I will make two assumptions. First, I will assume that the things that get explained by explanations – with a technical term, the *explananda* – are facts, and that these facts can be put into a contrastive form, that is, an "A rather than B" form. An explanation thus may explain why Tom is depressed rather than not being depressed; or why he is depressed rather than manic; or why he has been depressed since August rather than having been depressed for a longer or shorter time. Not much in the discussion will hinge on this assumption, but settling on one specific form of explanandum will increase both brevity and clarity. In addition, it has been made by many authors working on scientific explanation, from Van Fraassen (21) to Woodward (22).

When we start thinking about features of explanations that could stand in the way of their being combined, one rather trivial feature will come to us immediately: logical inconsistency. If I explain Tom's depression from that fact that he has been working too much and you explain it from the fact that he has been jobless, we are contradicting each other and no integration is possible. In order to avoid this, I will stipulate that in all the examples to be discussed later on, the explanations given are true; and furthermore, I assume – this is my second substantive assumption – that true statements are always logically compatible. Many will regard this assumption as a self-evident truth;

I myself do not; but I will assume its truth here in order to focus on the issues at hand.

Given this second assumption, there seems to be a strong presumption in favor of the idea that all explanations of the same explanandum will be combinable. After all, we can simply put them together; there being no logical incompatibility, nothing could stop us from doing so. This, I take it, is precisely why integrative approaches to science are so intuitively persuasive: if all our final theories are true, it surely *must* be possible to combine them. But of course, there are many ways to "combine" explanations, and it behooves us to take stock of them – and of any presuppositions they entail – before coming to a judgment about the matter.

In the following, I will identify three ways in which explanations can be combined: by presenting additive causes, by presenting different parts of a single causal tree, or by describing supervening levels. After a brief discussion of these three kinds of compatibility, I will argue that all of them share a basic presupposition that I will call *counterfactual compatibility*. This will suggest a way that even true, logically consistent explanations of the same fact can fail to be combinable: by *counterfactual incompatibility*.

As our example explanandum, let us take the fact F that patient P suffers from major depressive disorder (MDD), rather than not suffering from it. Let us postulate that P's MDD can be causally linked to a life history that has led to self-esteem and relationship issues; that the depression has been triggered by the loss of a job and the death of his best friend; that on a neural level the depressive symptoms are caused by, among other things, a disruption of neuroplasticity; and that P's self-esteem issues can be related to the exaggerated expectations his authoritarian father had of his only son. Given this situation, both of the following are acceptable explanations of F:

(1) P suffers from MDD because he lost his job.
(2) P suffers from MDD because his best friend died. These explanations both present causal factors that increased the likelihood of a depression and were in fact causally linked to it. Irrespective of whether either of them was sufficient for the occurrence of MDD, or whether both together were needed to trigger it, these causes can be added to each other in a single, more encompassing explanation:
(3) P suffers from MDD because he lost his job and his best friend died. This is what I call the *presentation of additive causes*: when two or more explanations present different causal factors that are independent but both increase the probability of the explanandum, we can simply combine them into a single conjunctive causal factor that is more informative than either of the factors alone. Of course, it is also possible that a set of explanations presents causal factors that are not independent, but that depend on each other because they are causally linked. Take, for instance, the following:
(4) P suffers from MDD because he lost his job and has been unable to find a new one.
(5) P suffers from MDD because he has self-esteem issues, which made him ineffective in his last job and caused him to lose it. The loss of his job triggered MDD.
(6) P suffers from MDD because the economy is in a slump and that has made him unable to find a new job. If he had found a new job soon after losing his last one, MDD would not have been triggered.

The relationship between these explanations is that each of them traces out a different part of a single causal *tree*, where a causal tree is the structure that is generated by providing the direct causes of one event, and then continuing to provide causes for any event in the tree whose causes have not been given yet. In this case, (4) explains F by giving two of its causes: the loss of the job and the inability to find a new one. (5) explains F by giving only one of those causes – the loss of the job – but by also explaining what caused that cause, thus moving up a level in the explanatory tree. Explanation (6) gives another of the causes of F – the inability to find a new job – and gives the causes of that cause. It is of course possible to combine (4–6) into a single, more complete description of the explanatory tree:

(7) P suffers from MDD because he has self-esteem issues and because the economy is in a slump. The self-esteem issues caused him to be ineffective at his last job, which in turn caused him to be fired. Because of the economic slump, he has been unable to find a new job. The prolonged joblessness was one of the things that triggered P's current episode of MDD.

This is what I call the *presentation of different parts of a causal tree*. Of course, the addition of causes and the presentation of different parts of a causal tree can be combined more or less *ad infinitum* in order to trace out the entire causal history of the event in the explanandum. Each of the explanations gives a different part of the tree, gives us a different set of events and causal links between them, and as this proceeds, we know about a larger part of the tree and understand the explanandum better.

These two ways of combining explanations are straightforward and important in practice, but they pose few theoretical problems. Things become more interesting when we move to two explanations like these:

(8) P suffers from MDD because he has a high stress level and stress causes the symptoms known as depression.
(9) P suffers from MDD because he has abnormal levels of cortisol, serotonin, and norepinephrine. These abnormal levels reduce the neuroplasticity of P's brain, which in turn causes the symptoms known as depression.[2]

We cannot understand (8) and (9) as tracing out different parts of a causal tree, for the simple reason that – at least on standard theories of the mental – they trace out the *same* part, but described at different levels or in different vocabularies. Where (8) speaks about stress, (9) speaks about the abnormal levels of

[2]For the potential relation between stress hormones, neuroplasticity, and depression, see Maletic et al. (23) and Pittenger and Duman (24).

certain hormones, but these are two descriptions of the same state. It is both possible and enlightening to combine (8) and (9):

(10) P suffers from MDD because he has a high stress level, which involves him having abnormal levels of cortisol, serotonin, and norepinephrine. These abnormal levels reduce the neuroplasticity of P's brain, which in turn causes the symptoms known as depression.

Philosophical questions about this situation remain, especially about the status of the word "involves" in (10). Is having stress *identical* to having certain hormonal levels, or does having stress instead *supervene*[3] on hormone levels? If it supervenes, could there be a reduction of theories about stress to theories about hormones, or are reductions impossible? Might it even be the case that this description of the situation is wrong, and that stress and certain hormone levels are merely accidentally cooccurring? Such questions are familiar from the philosophy of mind and will not be resolved any time soon. But for our current discussion, it turns out, perhaps surprisingly, that the answers to these questions make no difference. On any of the options in the debate, either (8) and (9) can be combined into (10) or at least one of them is false:

- On a reductionist theory, the two explanations are simply saying the same things in different vocabularies; once this is seen, the combination is trivial, because it turns out that there is nothing to combine.
- On a non-reductionist theory which sees psychological notions like "stress" as supervening on neurochemical states, the two explanations can both be given, and then linked through supervenience relations to result in a more complete explanation. This is what I will call the *description of supervening levels*.
- On a non-reductionist theory that rejects the supervenience thesis and instead believes that psychological events such as stress and neurochemical events such as high hormone levels are wholly distinct but related through the relation of causation, the two explanations can be combined by giving the causal interrelations between them. In this case, combining (8) and (9) into (10) turns out to be a case of *presentation of different parts of a causal tree*.
- On a radical dualist theory which sees mental events like stress and physical events like hormone levels as wholly distinct and non-interacting, explanation (9) must be false, for hormone levels cannot cause stress. So in this case, too, we do not have two true explanations of the same phenomenon that cannot be combined; we have a true and a false explanation, and the false explanation must be rejected.

[3]Supervenience is a notoriously difficult term to define adequately, but in this article, I will take it to be the relation such that (a) the values of supervening properties at time *t* are fully determined by the values at time *t* of the properties they supervene on, but (b) the supervening properties cannot be identified with the supervened-on properties. Many philosophers have defended the idea that while, say, mental states are not identical to brain states; nevertheless, our brain states fully determine our mental states. If this is so, then mental states supervene on brain states in the sense I am using the term here.

Which of these options is correct will be highly relevant to our view of the relation between psychology and neuroscience. But what anyone can seemingly agree on is that once we have found the true explanations, those explanations *can* be combined into a single story – either by identifying them, by linking them through supervenience relation, or by linking them through causal relations.

Having seen three important ways in which true explanations can be combined, and are combined in practice, we are still faced with the question of whether there are any conditions under which they cannot. To answer that question, we must think about what explanations are and how something could fail to be an explanation, even though its parts are explanations.

When we do think through the properties of explanations, we quickly find that they are not merely lists of unconnected facts. Explanations always trace links between the fact to be explained and other facts. Different theories of explanation have different ideas about what these links are like: according to Hempel's original DN-model, explanations show how the fact to be explained can be derived from other facts through laws of nature; according to unificationist theories, explanations show how the fact to be explained can be derived using unifying arguments; according to causal theories of explanation, explanations explain a fact by giving its causal antecedents [see Salmon (25) and Woodward (26) for overviews]. But what all these theories have in common, and what is indeed one of the central facts about explanation that any theory of explanation would have to do justice to, is that explanations allow us to draw *counterfactual conclusions* about the explanandum. To *know that* P suffers from MDD is to know something important; but to *understand why* P suffers from MDD is to have, in addition, a measure of insight into the conditions under which he would *not* have suffered. Explanations allow us to make claims about what *would have happened* in different circumstances. And this is indeed one of the prime reasons that we are interested in explanations at all, for by allowing us to see what would happen in different circumstances, they allows us to make an informed choice between different courses of action. [For more on the relation between explanation, causation, and counterfactuals, see Chapter 3 of Woodward (22).]

If one of the central obligations on an explanation is to allow us to draw counterfactual conclusions about the explanandum, then it is reasonable for us to require explanations to fulfill that obligation. To be precise, it is reasonable to ask of any explanation that the counterfactual consequences that follow from it are consistent: that is, that we cannot show from it both that if A had happened, C would have happened; and that if A had happened, C would not have happened. In other words, the counterfactual picture painted by any explanation should be coherent.

This in turn suggests a condition that two explanations of the same fact have to fulfill in order to be combinable into a single explanation: they should not have logically incompatible counterfactual consequences. If they do not, we will call them *counterfactually compatible*. If, on the other hand, they do have logically incompatible counterfactual consequences, we will call them *counterfactually incompatible*. The claim I am making, then, is that two true explanations of the same fact are combinable into one explanation only if they are counterfactually compatible.

(This is a necessary condition. Perhaps it is also sufficient, but I have no argument to that effect.)

In all our previous examples, the explanations were indeed counterfactually compatible. Both (4) and (5) imply that if P had not lost his job, he would not have had MDD. In addition, (5) implies that if P had not had self-esteem issues, he would not have lost his job; this is of course compatible with the previous claim. Both (4) and (6) imply that if P had been able to find a new job soon, he would not have had MDD. In the case of (8) and (9), the counterfactual implications are different but logically compatible: (8) implies that P would not have suffered from MDD if he had not suffered from stress, whereas (9) implies that P would not have suffered from MDD if his hormone levels had been normal; and these two claims are perfectly consistent on both reductive and non-reductive theories of the mental.

We must now ask ourselves whether it is ever possible for two true explanations to be counterfactually incompatible. Let us first look at an example involving two very different explanations of the same fact, one from the perspective of textbook physics and one from the perspective of common sense teleology. Suppose that a door in my living room is open rather than closed. Why? Here are two explanations:

(11) The door is open because a force greater than F was applied to it from the inside while the handle was down.
(12) The door is open to allow fresh air to get in.

Both of these explanations can be true at the same time. But now let us ask the following question: would this door have been open if it had been a door to the cellar instead of a door to the garden? The physicist, with (11) in hand, would say that, yes, the door would still have been open. After all, cellar doors do not have physical properties that make them physically more difficult to open than garden doors. But the common sense thinker, looking at (12), would say no, the door would have been closed if it had been a cellar door. After all, cellar doors are not opened to let in fresh air. Who of the two is right? Would this door have been open if it had been a door to the cellar? Well, yes *and* no – it depends on the perspective we are taking. But this means that the explanations from the two perspectives, while both valid and true, fail the test of counterfactual compatibility and cannot be combined into a single coherent explanation.

One might object that any incompatibility here is the result of the incompatibility of a broadly causal and a broadly teleological perspective; and one might then go on to claim that teleology has no place in science. If that is true, then examples like the one above could show at most that science cannot always be integrated with common sense; but this does not disprove the integrationist claim that the sciences themselves are always capable of being integrated. Perhaps this is true; although it would already be an interesting result, since discussions about teleology are by no means dead in science. But counterfactual incompatibility can in fact also arise between two perspectives that are both purely causal.

Let us return to our poor patient P, and let us ask the following question: suppose that P had been a woman, would he still have suffered from MDD? One way to approach this question – the approach that would be favored by a neuroscientist – would be to review the differences between male and female brains. Let us supposes that there is no systematic difference between the sexes such that female brains handle abnormal levels of cortisol, serotonin, and norepinephrine differently from male brains. Then, the neuroscientist would pronounce, correctly and with ample justification, that if P had been female, (s)he would still have suffered from MDD.

But the question could also be answered by P's therapist, who has been especially interested in talking through his life history with him, with a special emphasis on traumatic events from his early childhood. According to this therapist the crucial cause of P's self-esteem issues is the way P's father treated his only son; a way that was markedly different from the way he treated his daughters. The therapist thus comes to the conclusion – just as correct and just as justified as that reached by neuroscientist – that if P had been a woman, (s)he would not have suffered from MDD.

There is nothing especially mysterious about this situation. Different scientific perspectives on P naturally lead to different ways of evaluating counterfactual claims about him. For a neuroscientist, contemplating the influence of gender means contemplating the way that gender has influenced the structure and functioning of the patient's brain. For the therapist, contemplating the influence of gender means contemplating the way that gender has influenced the patient's life history. Both of these perspectives are equally valid, and both lead to explanations that should be accepted. But these explanations cannot be accepted into one single coherent explanation; for combining them leads to a story in which P would both have suffered from MDD and not suffered from MDD if he had been a woman. So, the therapist's life-history approach and the neuroscientist's approach have to remain isolated to a certain extent. Here, we have a case of counterfactual incompatibility; and in general, counterfactual incompatibility may occur when we try to integrate explanations from different research approaches. When it does, it acts as a barrier to integrative pluralism.

This conclusion could be attacked in two ways. First, one could attack the claim that counterfactual compatibility is a requirement for two explanations to be combined. Now, admittedly, by choosing a suitable low standard for what "integration" means, one can always claim that two research approaches can be integrated. But counterfactual incompatibility is a real barrier to any substantive kind of integration, because it means that we cannot simply transfer conclusion reached in one approach to the other approach. If the neuroscientist finds that gender is irrelevant to MDD, the therapist or the sociologist cannot just accept that conclusion; for the conclusion, while true – in our example – from the neuroscientific perspective, might well be false from the other perspectives. This non-transferability of counterfactual conclusions is surely a good reason to hold that the different research approaches are to some extent isolated and autonomous.

Second, one could claim that, my examples notwithstanding, counterfactual compatibility cannot occur between true explanations. For, one could argue, it is logically impossible that "if A had happened, then B would have happened" and "if A had happened, then B would not have happened" are both true. To substantiate this conclusion, one could appeal to influential theories about the

truth conditions of counterfactuals. Lewis (27), for instance, tells us that "if A had happened, then B would have happened" is true just in case that B is true in the closest possible world where A is true, where the closeness of possible worlds is defined in terms of their similarity to ours. If such a story were correct, and, crucially, *if similarity were a non-contextual affair*, something that should be evaluated in the same way across all the sciences, then either the therapist or the neuroscientist would have to be wrong. To see which, we would have to find out which world is more similar to ours, the one envisaged by the therapist or the one envisaged by the neuroscientist. And whichever of them in their imaginative flights stayed closer to home, so to speak, would be the person drawing the correct counterfactual conclusions.

Such a procedure, however, has very little to recommend itself. Theories about the truth conditions of counterfactuals should respect our everyday evaluations of counterfactuals; and it is an undeniable fact that people working from different perspectives use different scenarios to evaluate the same counterfactual claims. As Lowe (28) points out, the truth conditions of counterfactuals are highly context-dependent. Lowe then argues (pp. 54–55) that the context influences how we evaluate claims about the similarity of possible worlds, and that this context is at least partly defined by the intentions of the speaker. For our current purposes, we can slightly modify his proposal and state that the context within which counterfactuals are evaluated is at least partly defined by the research approach within which the claim appears. Counterfactual evaluation in neuroscience takes scenarios into account that are ignored in the therapeutic setting, and the other way around. The different sciences use different relevance criteria; and this does not make a difference not only for which facts they uncover but also for how they reason about counterfactual scenarios. Since explanations are tightly connected to counterfactual scenarios, these differences between research approaches translate into an incompatibility of the explanations they generate.

This concludes my argument for the claim that I set out to prove, namely, that true explanations of the same fact sometimes cannot be combined into a single bigger explanations. Counterfactual incompatibility is a barrier to such combination, and counterfactual incompatibility is real. This result nicely mirrors that of Lange (29). His point is that different research approaches take different sets of counterfactuals seriously, and that this leads to incompatible laws; my point is that different research approaches sometimes reach incompatible results when evaluating identical counterfactuals. Both points support the conclusion that research approaches can be expected to be at least partly autonomous.

Let me end this section by professing ignorance about two points. First, I am not sure whether counterfactual incompatibility can also occur within a single research approach – e.g., whether two true neurological explanations of a brain phenomenon could ever turn out to be incompatible. If this were possible, science would be even more disunified than we tend to think. Second, I do not know whether this section has covered all the ways in which explanations can be combinable or fail to be combinable. In this respect, I make no claim to having exhausted the territory.

CONCLUSION

My analysis of the debate surrounding the unity of science and explanatory pluralism revealed that the most pressing issue lies in the choice between integrative and isolationist pluralism; or rather, in finding out whether the integrative tendencies present in current science should be fully indulged in, or should be held in check by affirming that a certain amount of isolation is unavoidable. I further distilled this issue into the question of whether two true explanations of the same fact could ever fail to be combinable into one single explanation. It turns out that although many explanations are in fact combinable, this only holds when they have compatible counterfactual consequences. I then argued that true explanations from different sciences can have incompatible counterfactual consequences. This leads us to the general conclusion that a certain amount of isolation between the sciences is indeed both present and unavoidable; forcing all the sciences to use the counterfactual relevance criteria of one of them would rob us of part of the insight that the different sciences can give us and would lead to the uncritical transfer of counterfactual claims from one science into another, with potentially disastrous results (in the case of, e.g., a sociologist who rejects the possibility that gender could be related to psychological conditions because the neuroscientists tell him that there is no such relation). This does not mean that we should not strive for integration and the benefits of interdisciplinarity. But it does mean that we should not take this project too far, for, at the end of the day, there will still be the different sciences with their different perspectives and insights. The plurality of the sciences is to be cherished rather than combated.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

1. Cat J. The unity of science. Winter 2014 ed. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy* (2014). Available from: http://plato.stanford.edu/archives/win2014/entries/scientific-unity/
2. Pombo O, Symons J, Torres JM. Neurath and the unity of science: an introduction. In: Symons J, Pombo O, Torres JM, editors. *Otto Neurath and the Unity of Science*. Dordrecht: Springer (2011). p. 1–11.
3. Oppenheim P, Putnam H. The unity of science as a working hypothesis. In: Feigl H, Scriven M, Maxwell G, editors. *Minnesota Studies in the Philosophy of Science* (Vol. **2**), Minneapolis: Minnesota University Press (1958).
4. Fodor J. Special sciences: or the disunity of science as a working hypothesis. *Synthese* (1974) **28**:97–115. doi:10.1007/BF00485230
5. Dupré J. *The Disorder of Things. Metaphysical Foundations of the Disunity of Science*. Cambridge, MA: Harvard University Press (1993).
6. Cartwright N. *The Dappled World. A Study of the Boundaries of Science*. Cambridge: Cambridge University Press (1999).
7. Teller P. Twilight of the perfect model model. *Erkenntnis* (2001) **55**:393–415. doi:10.1023/A:1013349314515
8. Ladyman J, Ross D. *Every Thing Must Go. Metaphysics Naturalized*. Oxford: Oxford University Press (2007).
9. Symons J, Pombo O, Torres JM, editors. *Otto Neurath and the Unity of Science*. Dordrecht: Springer (2011).
10. McCauley RN, Bechtel W. Explanatory pluralism and heuristic identity theory. *Theory Psychol* (2001) **11**:736–60. doi:10.1177/0959354301116002

11. Kendler KS. Toward a philosophical structure for psychiatry. *Am J Psychiatry* (2005) **162**:433–40. doi:10.1176/appi.ajp.162.3.433
12. Abney DH, Dale R, Yoshimi J, Kello CT, Tylén K, Fusaroli R. Joint perceptual decision-making: a case study in explanatory pluralism. *Front Psychol* (2014) **5**:330. doi:10.3389/fpsyg.2014.00330
13. Marchionni C. Explanatory pluralism and complementarity: from autonomy to integration. *Philos Soc Sci* (2008) **38**:314–33. doi:10.1177/0048393108319399
14. Mitchell S. *Unsimple Truths. Science, Complexity, and Policy*. Chicago: University of Chicago Press (2009).
15. Campaner R. Explanatory pluralism in psychiatry: what are we pluralists about, and why? In: Galavotti MC, et al., editors. *New Directions in the Philosophy of Science*. Cham: Springer (2014). p. 87–103.
16. Van Bouwel J. Pluralists about pluralism? Different versions of explanatory pluralism in psychiatry. In: Galavotti MC, et al., editors. *New Directions in the Philosophy of Science*. Cham: Springer (2014). p. 105–19.
17. Brigandt I. Beyond reduction and pluralism: toward an epistemology of explanatory integration in biology. *Erkenntnis*. (2010) **73**:295–311. doi:10.1007/s10670-010-9233-3
18. Andler D. Unity without myths. In: Symons J, Pombo O, Torres JM, editors. *Otto Neurath and the Unity of Science*. Dordrecht: Springer (2011). p. 129–44.
19. Van Bouwel J, Weber E, De Vreese L. Indispensability arguments in favour of reductive explanations. *J Gen Philos Sci* (2011) **42**:33–46. doi:10.1007/s10838-011-9141-5
20. Gervais R. A framework for inter-level explanations: outlines for a new explanatory pluralism. *Stud Hist Philos Sci* (2014) **48**:1–9. doi:10.1016/j.shpsa.2014.07.002
21. Van Fraassen B. *The Scientific Image*. Oxford: Oxford University Press (1980).
22. Woodward J. *Making Things Happen*. Oxford: Oxford University Press (2003).
23. Maletic V, Robinson M, Oakes T, Iyengar S, Ball SG, Russell J. Neurobiology of depression: an integrated view of key findings. *Int J Clin Pract* (2007) **61**:2030–40. doi:10.1111/j.1742-1241.2007.01602.x
24. Pittenger C, Duman RS. Stress, depression, and neuroplasticity: a convergence of mechanisms. *Neuropsychopharmacology* (2008) **33**:88–109. doi:10.1038/sj.npp.1301574
25. Salmon W. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press (1989).
26. Woodward J. Scientific explanation. Winter 2014 ed. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy* (2014). Available from: http://plato.stanford.edu/archives/win2014/entries/scientific-explanation/
27. Lewis D. *Counterfactuals*. Oxford; Cambridge, MA: Blackwell Publishers; Harvard University Press (1973).
28. Lowe EJ. The truth about counterfactuals. *Philos Q* (1995) **45**:41–59. doi:10.2307/2219847
29. Lange M. Who's afraid of *Ceteris-Paribus* laws? Or: how I learned to stop worrying and love them. *Erkenntnis* (2002) **57**:407–23. doi:10.1023/A:1021546731582

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 16 February 2016
doi: 10.3389/fpsyt.2016.00012

A Reconciliation for the Future of Psychiatry: Both Folk Psychology and Cognitive Science

*Daniel D. Hutto**

Faculty of Law, Humanities and the Arts, School of Humanities and Social Inquiry, University of Wollongong, Wollongong, NSW, Australia

Philosophy of psychiatry faces a tough choice between two competing ways of understanding mental disorders. The folk psychology (FP) view puts our everyday normative conceptual scheme in the driver's seat – on the assumption that it, and it only, tells us what mental disorders are (1). Opposing this, the scientific image (SI) view (2, 3) holds that our understanding of mental disorders must come, wholly and solely, from the sciences of the mind, unfettered by FP. This paper argues that the FP view is problematic because it is too limited: there is more to the mind than FP allows; hence, we must look beyond FP for properly deep and illuminating explanations of mental disorders. SI promises just this. But when cast in its standard cognitivist formulations, SI is unnecessarily and unjustifiably neurocentric. After rejecting both the FP view, in its pure form, and SI view, in its popular cognitivist renderings, this paper concludes that a more liberal version of SI can accommodate what is best in both views – once SI is so formulated and the FP view properly edited and significantly revised, the two views can be reconciled and combined to provide a sound philosophical basis for a future psychiatry.

Keywords: philosophy of mind, narrative therapy, cognitive neuroscience, philosophy of psychiatry, philosophy of cognitive science

OPEN ACCESS

Edited by:
Derek Strijbos, Radboud University, Netherlands

Reviewed by:
Philip Gerrans, University of Adelaide, Australia
Rachel Valerie Cooper, Lancaster University, UK

***Correspondence:**
Daniel D. Hutto
ddhutto@uow.edu.au

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry

Received: *17 November 2015*
Accepted: *23 January 2016*
Published: *16 February 2016*

Citation:
Hutto DD (2016) A Reconciliation for the Future of Psychiatry: Both Folk Psychology and Cognitive Science. Front. Psychiatry 7:12. doi: 10.3389/fpsyt.2016.00012

> There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy.
> *Hamlet* Act I, Scene 5, 167–8

FOLK PSYCHOLOGY RULES

How should we best understand, categorize, and treat mental disorders? A familiar answer in philosophical circles is that any approach to mental health must always operate with reference to the normative features that define our folk psychological (FP) understanding of mind. Call this the FP view of mental disorders and psychiatry. Its driving assumptions are that FP, and only FP, conceptually defines what it is to have a mind because FP, and FP alone, supplies the necessary and sufficient mark of the mental, emphasizing its essentially rational character. On the standard, narrow reading, FP plays this governing role precisely because it is understood to be the commonsense theory or conceptual scheme that reveals how mental states – typically assumed to be propositional attitudes – interact in the rational production of behavior and action.

Graham (1), a staunch spokesperson for the FP view, advances a theory of mental disorder according to which we have no choice but to make reference to reason and rationality when understanding

such phenomena because such features "help to constitute and define distinctively mental activity such as believing, hoping, desiring, deciding, thinking and the like" [(1), p. 7].

The main idea behind this vision is that FP supplies the only normative standard of what minds are and how they operate. Hence, only against FP's standard is it even possible to detect and demarcate mental disorders. Mental disorders arise when things go awry with us at some level, when in some important sense, a person fails to live up to or systematically violates the standards of rationality that characterize our everyday folk psychological ways of thinking. There may be various mental and non-mental causes of such failures. But, simply put, FP is the necessary reference point of what a normatively defined well-functioning mind looks like and what its rational characteristics are. Disordered minds, by comparison, are less than flourishing minds that fail to meet that normative standard.

Even though proponents of the FP view accept that perfect rationality is a notoriously slippery notion that evades precise analysis they insist, nonetheless, that, "rationality is essential to mindedness" [(1), p. 12]. Moreover, it is assumed that rationality is only something exhibited by whole persons and not by the operations of their subpersonal parts. Putting all of this together, in standard formulations, the FP view holds that we have no choice but to understand minds by making use of FP concepts, which apply to persons whose intentional attitudes exhibit an inherent rationality.

> Thinking of intentional attitudes (beliefs, desires, and so on) as presupposing the rationality of persons makes it clear that we persons are purposive or goal directed in behavior and that how we act or behave depends on our purposes or reasons for acting [(1), p. 120].

Graham (1) dubs this the rationality-in-intentionality (RIT) thesis for short. RIT takes rationality to be the hallmark of the mental – one that sets the mental forever apart from all other kinds of phenomena, and this is what makes the mental irreducibly autonomous.[1] The autonomy of the mental thesis can be understood in more or less realistic terms. Yet in all versions the root idea, subscribed to by all fans of the FP view, is this: propositional attitudes can only be ascribed, or only have life, when they stand in appropriate kinds of holistically and normatively defined rational relations. Mental phenomena exist if and only if the relevant forms of rationality are in place: viz. they live in the space of reasons. This is allegedly why when rationality is absent, we must switch to another scheme for understanding the relevant phenomena; in such cases, a move to non-mental concepts and explanatory schemes becomes necessary precisely because minds, properly understood, are fading or absent.

The crucial assumption of the FP view is that it is the job of philosophy of mind to reveal and articulate the essential contours of our commonsense understanding of the mind – which are assumed to be *the only* bona fide conception of mind.[2] The standard view is that this can only be achieved by means of some kind of conceptual analysis or radical interpretation [see, e.g., Ref. (5)].

With these assumptions about the essential characteristics of minds in place, the FP view of mental disorders firmly opposes what, by its lights, is its only possible rival, a scientifically orientated, non-mentalistic FP-eliminativist approach – one that looks solely to neuroscience to discern "the best understanding of and treatment for mental disorder" [(1), p. 6]. A purely brain-based approach to mental disorders is oxymoronic from the perspective of those who hold that FP defines the mental; such an approach might tell us much about non-mental disorders of various kinds, but it could not be a starting point of inquiry into psychiatry because it misses out the mental altogether. Despite insisting on this point, fans of the FP view do not deny that neuroscience can play a part in the larger business of psychiatry. They do insist, however, that the part the brain sciences can play is always and everywhere secondary, servile, and subservient. Crucially, the FP view of psychiatry "does not relinquish the theory (of mental disorders) to, but deploys, brain science" [(1), p. 9].

Although clearly incompatible with a purely scientific, eliminativist vision of psychiatry, the FP view is compatible with making explanatory use of a range of scientific findings. The sciences of the mind have something important to add to the story so long as they take their direction from FP when it comes to understanding and classifying mental disorders on the basis of possible causes. There is no contradiction to be found in such a cooperative enterprise for those who think FP defines the mind: but this is so only as long as it is accepted that FP must always remain in the driver's seat when coordinating any such combined efforts [(1), p. 11].[3]

The logic is straightforward. Mental disorders – on the FP view – only ever show up as disturbances within the space of reasons. Even so, there can be non-mental causes of mental disorders. We can think of such causes as arational, non-mental disorder influences that, to use Graham's apt phrase, "gum up" "the rational works" [(1), p. 160]. Non-mental factors – the influences of brain and behavior – can interfere with and upset our rationally constrained mentality. Accordingly, non-mental factors can contribute to and help explain the occurrence of mental disorders – and this can happen even if the non-mental mechanisms in question are in perfect order and are operating just as they should.[4]

All in all, the FP view insists on a particular understanding of the explanatory relations that can hold between the mental and the

[1] The FP view endorses the irreducibility claim about the mind, which Davidson championed long ago: "The reason mental concepts cannot be reduced to physical concepts is the *normative* character of mental concepts" [(4), p. 46].

[2] Those who take the autonomy thesis seriously – whether in strongly realist or more interpretationist renderings – maintain that an FP understanding of mind plays an absolutely essential role in defining what minds are and, hence derivatively, what *mental* disorders are. Accordingly, "the mind *qua* mind puts its inscription on the sources of a disorder. We cannot recognize a mental disorder without uncovering that mark" [(1), p. 11]. It is because of the need for benchmarking against the mark of the mental – which can only be done *via* FP – that "no conceptually regimented and normatively informed theory of mental disorder can be devised without taking philosophy of mind seriously" [(1), p. 1].

[3] Thus in line with this, Graham (1) argues that going the FP-governed way ought not to encourage one to endorse the DSM atheoretical method of classifying and characterizing mental disorders.

[4] For example, Graham (1) illustrates the latter point vividly by reminding us that, "an addict … does not possess a broken brain" (p. 179). The reason this can be so, Graham (1) claims, is because, "The brain, in general, is not hard-wired for personal prudence. Neural activity may systematically underwrite unwise behaviors without exemplifying a breakdown or something wrong or damaged in its wetware or machinery" (p. 178).

non-mental when it comes to making sense of mental disorders. Neuroscience can help us to understand, for example, the condition of the unwilling addict because reference to brute arational neural mechanisms can help "*to explain* why addicts suffer from relapse in spite of themselves" [(1), p. 179, emphasis added]. Non-mental mechanisms – whether malfunctioning or not – feature in the larger story of specific mental disorders because they may explain what interferes with the rational works. Telling the right story about such non-mental influences is complicated by the fact that there can be "different hypotheses about the irruptive role of a-rational neurobiological/neurochemical mechanisms into the space of reasons" [(1), p. 178].

Speaking on behalf of the FP vision of psychiatry, Graham (1) sees no difficulty in asserting that, "even though mental disorders are not brain disorders, neuroscience helps to illuminate *their nature*" [(1), p. 13, emphasis added]. How should we understand this claim? It deserves attention for, as just noted above, the FP view holds that the essence of *mental* disorders can only be understood with reference to what occurs in the space of reasons. Accordingly, explanations that cite non-mental goings-on from outside that space can only shine a light on the nature of mental disorders if we distinguish the *essential* characteristics of the latter from explanations that tell a fuller story about their *actual* natures (or, more likely, how non-mental mechanisms make an actual but effective difference in particular cases).

This requires drawing a distinction between the *essence* of mental disorders and their *actual* natures. A standard way to do this is to take a leaf out of the analytic playbook of commonsense functionalism (6–8). Analytic functionalists hold that our everyday folk theories or practices conceptually define what minds are, fully and completely. On this score, the so-called sciences of the mind reveal nothing about the essential features of minds. *Prima facie*, this may seem extreme. However, adopting this view is entirely compatible with the idea that the sciences can discover much about how mental phenomena are, *as a matter of fact*, instantiated in the actual world – hence, the sciences can discover much about their *actual nature*. They can do so once FP tells "empirical inquiry what to look for" [(2), p. 51]. If we distinguish the essential and actual characteristics of mental phenomena and, relatedly, mental disorders, it becomes possible to understand how the sciences can assist with an understanding of the nature of mental disorders on the FP view. *Pace* Locke, on this vision, it is the scientists of mind, not the philosophers of mind, who must do the under-laboring.

PSYCHIATRY IN THE SCIENTIFIC IMAGE

The FP view of psychiatry has some fierce critics. It has been accused of presenting an unjustifiably restrictive vision of the role the sciences of mind play in mental health. Advancing the idea that psychiatry needs to be grounded entirely in the cognitive sciences, Murphy (2) is the foremost defender of the scientific image (SI) view of psychiatry. He and his friends see the FP view as unacceptable because it denies the sciences of the mind a free hand in revealing the essential character of mental phenomena. That restriction, SIers hold, results in an unwarranted fettering of psychiatric explanations and classifications [(2), p. 48, 51].

Adherents of the SI view are self-styled progressives. They insist that psychiatric explanation and nosology should not be regimented by, or beholden to, commonsense intuitions or assumptions. Psychiatric explanations, they hold, require no guidance, warrant, or mandate from FP. Proponents of the SI view insist that the future of the mental health field depends on fully embracing the sciences of the mind. The core assumption is that the sciences provide the requisite tools for a free inquiry into the nature of mental disorders. Moreover, the scientific work is to be conducted without requiring any appeal to commonsense notions of the mind, as filtered and understood through philosophy.

In place of the FP vision of psychopathology, Murphy (2) sets out his stall for a revisionary objectivism about the nature of mind. Accordingly, the bid is to discover what minds are through the development of a pragmatic and open-ended, scientifically driven conceptual framework, one that is revisable in practice and one that rests on testing out a series of empirical bets about mental phenomena.[5] Crucially, the scientific investigations Murphy envisages would not be shackled by FP's oversight. On the SI vision, to truly explain and classify normal and abnormal minds, we must look to our best cognitive sciences and to those alone.[6]

An immediate consequence of embracing the SI view is that it open up the scope of what we might think of as the mental and how we might think of it, quite considerably. By implication, the same goes for mental disorders. For example, given that perception is a paradigmatically mental phenomenon, it turns out that, on the SI view, blindness – however counterintuitive it may seem – counts not just as a disorder of the visual system but as a *mental* disorder [(2), p. 54, 55–57]. Murphy is happy to bite this and other, similar bullets. The justification is simple: violating a few folk intuitions is a small price to pay if going the purely SI way puts psychiatry on a "sounder footing" [(2), p. 11].

This is all very well and good as a sort of SI position statement but what justifies taking the SI path? Why suppose that SI might reveal new and deeper facts about how minds operate and how they can go wrong?[7] Why not hold that the sciences of the mind do whatever good work they do by functioning in exactly the way the FP view says they do – viz. by supplying "the empirical application of our pre-theoretic folk concepts" [(2), p. 50]?

The most straightforward and compelling answer is that there is surely more to the mental than dreamed of by FP. This conclusion is hard, if not impossible, to resist if FP characterizes the mental wholly in terms of the propositional attitudes and how

[5] In pressing for this future vision of mental health as wholly grounded in science, Murphy (2) laments that much contemporary psychiatry actively shies away from theory (as exemplified in the avowed theory-neutrality of the DSMs). Against the diagnostic descriptivism of the DSMs, he maintains that psychiatry must aim to find out how things "really are" with mental disorders and what underlies them. The only way psychiatry can do this, and thus secure its future, would be to take advantage of and actively contribute to developments in the cognitive sciences.

[6] A psychiatry cast in the scientific image must assume that, "what counts as normal human nature is decided by a variety of disciplines that comprise the cognitive and biological sciences" [(2), p. 11].

[7] Murphy (2) is certainly right that there is a real and urgent need for psychiatry to address the issues that lie at the heart of the debate between FP and SI views. For the troubling fact is that, "psychiatry as it stands is not a particularly mature or successful enterprise" [(2), p. 10].

they rationally inter-relate. For on any such a rendering, there is every reason to believe that

> mind has an existence and substantive character that *goes well beyond*, and is independent of our best common-sense interpretative practices. Hence knowing the truth about the mind requires *a great deal more* than informed reflection on those practices. In fact, it requires cognitive science [(9), xiv, emphases added].

A full understanding of all that is mental cannot be limited to FP characterizations alone. There are many aspects of mind – even quite ordinary, everyday ones – such as the complex ways that perceiving and acting interact – upon which FP, as construed above, has simply nothing to say. Such examples multiply. There are many forms and aspects of mentality that can only be understood by engaging in modes of inquiry that go beyond interrogating FP as traditionally conceived. FP casts no light on the properties and dynamics of basic minds [for an extensive discussion, see Ref. (10)].

Call this the "More to the Mind than FP" objection. It strikes at the core assumptions of the official FP view. Notably, however, even if the "More to the Mind than FP" objection defeats or should make us suspicious about the FP view, it does not, by itself, justify adopting the pure SI vision. For even if there is "More to the Mind than FP," it does not follow that our understanding of minds, and by implication mental disorders, can *only* and *wholly* be supplied by the sciences of the mind.

In the end, because FP is not the whole story about minds, it will be argued that going the SI way is best – but only if SI is carefully qualified. Why so? Because FP is part of the story of the mental: arguably, important aspects of human minds can only be understood in FP terms. This can be so even if FP assumptions about the mind should not be the basis for or otherwise restrict our investigations into the fundamental nature of minds. Before attempting to show how to marry these ideas in Section "Keeping FP in the Picture," the next two sections raise important doubts about standard cognitivist formulations of the SI view and their inherent neurocentrism.

A CERTAIN IRONY

The SI view is open to understanding the mind in new ways that go beyond FP. Despite the essential openness of the SI view, some of its most prominent proponents have tried to foreclose on certain possibilities. Based on assumptions about what the best explanations in the cognitive sciences will look like, some campaign for a neuro-based cognitivist version of the SI view. For example, under SI's auspices, Murphy offers a defense of the idea that, "psychiatry is a branch of medicine dedicated to uncovering the *neurological basis* of disease entities" [(2), p. 10, emphasis added].[8] For him, going the SI way paves the way for adopting the medical model of psychiatry such that the work of psychiatry becomes that of tracing "abnormalities in behavior and cognition to specific *causal factors that are realized in brain* tissue" [(2), p. 13, emphasis added].

In Murphy's mind, adoption of the SI view leads naturally to firmly recommending a merger of psychiatry and clinical neuropsychology.[9] He is supremely confident that a purely neuro-based approach will dominate the future of psychiatry. This is evinced by his commitment to neurocomputationalism, input–output functionalism, modules, and so on. But why assume that the brain's the thing? Retort: Who seriously doubts it? Murphy tells us that, "After all, *everyone knows that* psychological phenomena, like all human behavior, are *rooted in* brain processes" [(2), p. 9, emphasis added]. How should we interpret the "everyone knows that" operator in this statement and what epistemic backing does it have? There are several possibilities.

The first is to go "Folk Analytic." Perhaps we can appeal to folk intuitions about the mind to justify talk about what everyone knows about the brain basis of minds. Clearly, this is a non-starter for SIers. The SI view precludes making any appeal to hypothesized folk theories and the intuitions they sponsor in order to explain the epistemic credentials of "everyone knows" talk. Folk intuitions can give no backing to SI friendly claims about what everyone knows; hence, they cannot help justify the claim that cognition is wholly caused by and realized in the brain as opposed to having a wider and non-exclusively neural basis. Put simply, to adopt the SI view of psychiatry is to forego making appeals to folk intuitions in order to defend claims about "what everyone knows" about the mental. Call that Murphy's Law. Murphy too must abide by it.

The second way to go might be to appeal to consensus in this matter in the philosophy of cognitive science. Classical cognitivist approaches to cognition promote a brainbound account of cognitive processes by adopting a representationalist and internalist account of the vehicles of cognition. If everyone in the field agrees that cognition is always and everywhere content involving and that the vehicles of mental content are neural, then this would justify claiming that "everyone knows" neurocentrism to be true. The best cognitive explanations of behavior can, in effect, "throw away the world" and focus solely, and solipsistically, on the properties that supervene on the current internal neural states of cognizers (11–14).

SIers would be justified in saying that psychiatry ought to take an exclusive interest in brains if classical cognitivism were true. The trouble, for Murphy and followers, is that classical cognitivism may not be true, and – as things stand – it is far from a safe bet to assume that it is. More to the point, looking at the state of the philosophy of cognitive science, it can be safely said that classical cognitivism is *not known* to be true. There are deep-seated,

[8] Murphy's (2) defense is admittedly qualified because he admits there are limits to our understanding when it comes to naturalizing and mechanizing central reasoning processes such that it may turn out that a proper scientific understanding of the latter might never be attainable.

[9] Such a merger, he holds, is "necessary to develop the broadest and most fertile approach to understanding psychopathology" [(2), p. 12]. In saying this, he does not promote a crude reductionism. He does not assume that neuropsychology offers molecular explanations or that its explanations are somehow more fundamental. Nevertheless, he holds, neuropsychological explanations have a privileged status: they provide a special understanding that affords unique possibilities for intervening upon and treating mental disorders.

on-going philosophical debates about the character of cognition and the reach of cognitive processes – and these debates are far from being conclusively settled.

Murphy is well aware of these debates and their import. His official word on the matter in 2006 was to note that, "Some pictures of the mind stress embodiment very heavily … Others prescind from details of our embodiment to stress a more purely computational theory of the mental … what counts as the mental depends in part on who is right in these debates" [(2), p. 64]. Despite this acknowledgment, Murphy thinks there is really no doubt that the sciences of mind will stick to providing explanations in terms of brain-based, semantic representations. This he takes to be a settled issue – even if, in the end, the hypothesized brain-based representations in question turn out not to have contentful properties of the FP sort.[10]

However, what exactly are the defining properties of representations, as defined solely by the sciences of the mind, without any reference to the kinds of content understood by FP? How should we understand the disagreements between representationalist and non-representationalist if we do not appeal to some notion of content as supplied by FP or some other agreed upon non-FP theory (15)? And without agreement about the defining properties of representations understood in non-FP terms – which might be supplied if we had a well-developed non-FP theory of content – how are we to decide where the boundaries of mind and cognition lie? How are we to determine whether – in the end – the best explanations of sciences of the mind will be given in terms of "inputs" and "outputs" that are purely neural as opposed to involving extraneural factors too (16)?

Against this backdrop, Murphy's confidence in an exclusively representationalist and neuro-focused future for psychiatry will seem, at best, premature – and at worst, it will look like a groundless pledge of allegiance. For the fact is there is no agreement in the philosophy of cognitive science that supports the idea that everyone knows – at least, not yet – that psychological phenomena are rooted or realized exclusively in brain processes.

But wait. Surely, we are looking for consensus in the wrong place. The fact that philosophers – of the mind or otherwise – disagree about important topics is hardly news. Perhaps there is yet another, more properly scientific consensus that we can appeal to in order to make good on the "everyone knows" claim. Doesn't a quick glance at the current agreement in the theoretical commitments of actual scientists of the mind secure its truth? The great bulk of scientists of the mind do talk of neural and mental representations in free and easy ways these days. Does it follow that they are committed to a cognitivist take on mental representations of the sort described above – one that would justify neurocentrism? Establishing that would require serious and detailed interpretative work: it would need to be shown that the representational talk of scientists has all of the relevant commitments and that it is more than nominally unified. It is far from obvious that this is the case. One major problem is that no unified theory of representation currently exists. Worse still, if we look at the current state of cognitive science there does not seem to be a single, settled story to tell about which theoretical tools – representational or non-representational – are primary or the best ones to use when it comes to understanding cognition and explaining intelligent activity. We seem to be living in a mixed economy. If this is right, then there is not an existing scientific consensus SIers can point to in order to justify the claim that "everyone knows" cognition to be brainbound.

On top of this, even if such a current consensus did exist – even if all good cognitive scientists turned out to be representationalists in the relevant sense – more work would be needed in order to determine whether the entities and properties they posit now will stand the test of time. It is always possible that even if all cognitive scientists are currently committed to neural representations still, it might turn out that something with different properties will best explain the relevant phenomena. Cognitive science is, after all, an unfinished business. Hence, even if today's scientists did have common commitments that would justify adopting neurocentrism, we might still worry that any such contingent fact would not provide a secure basis for making firm predictions about the future of psychiatry. The official story is that not long ago cognitivism replaced old school behaviorism, right? Science is shifty – but in a good way. The SI view should surely embrace that.

At this stage of the game, there seems to be no obvious justification for fans of the SI view to reject the idea that psychiatry might look beyond the brain when it comes to understanding, explaining, and treating psychopathological disorders. Indeed, there are positive reasons for thinking that it is fruitful to look beyond the brain when it comes to understanding mental phenomena (16). Notably, since looking beyond the brain does not entail ignoring the brain, adopting such a liberal SI view is perfectly in line with a modified version of Murphy's assertion that "we are animals with a biology including a brain that is [part of] the foundation of our mental life" [(2), p. 10].

Still, it might be thought that the foregoing liberal assessment is too blithe, quick, and programmatic. Aren't there good grounds for thinking that cognitive science will remain deeply committed to cognitivism and neurocentrism, even if in the final reckoning, it deviates in some matters of detail from the classical versions of those views? Aren't there special reasons for favoring cognitivism – reasons that we can identify here and now – that would justify neurocentrism and thus rule out more radical and extensive possibilities for understanding the nature and extent of cognition. That seems to be the line of several prominent defenders of cognitivist variants of the SI view (2, 3).

COGNITIVE BRIDGE WORK?

In defending their predictions about the rightful dominance of an SI-based medical model, Murphy and Smart (17) make it clear that this future is to be secured by *cognitive* neuroscience

[10]Thus, in a forthcoming paper, Murphy writes, "The question whether science makes use of representational systems isn't really open to doubt any longer: many areas of psychology and neuroscience take for granted the existence of semantic interpretations of internal states of some cognitive system. The assumption that inputs and outputs to and from components of the brain represent distal features of the world has been part of neuroscience since the nineteenth century. What is open to doubt is whether representation, as used in the sciences of mind, has the properties that philosophers have found in intentional content, as presupposed by folk psychology. I am not taking a stand on that … " (Murphy D. *Brains and Beliefs* (Unpublished)).

and not merely some brutal brain science. What makes cognitive neuroscience special, they maintain, is that it posits subpersonal information-processing mechanisms that are at once both causal–mechanical and intentional in character. It is because it blends the cognitive with the neural that cognitive neuroscience has unique explanatory power: it, alone, allows for an integrated scientific story to be told about minds.

Looking exclusively at what goes on in brains is apparently justified because of the depth and unity cognitive neuroscientific explanations can provide. Gerrans (3) makes this case in great detail.[11] He argues that cognitive neuroscience understands, "persons as complex, hierarchically-organized information-processing systems implemented in neural wetware" [(3), p. 16].[12] Seeing persons as brain based, in turn, allegedly confers peculiar advantages because it puts us in a position, for example, to "show how *facts* identified and explained by disciplines operating at 'levels' such as molecular neurobiology or neuroanatomy *can explain* psychological and phenomenological level *facts* that give delusion its clinical profile" [(3), p. 20, emphases added].[13] What makes having a cognitive theory pitched at the subpersonal information-processing level so uniquely valuable is that it is needed to "*bridge the gap* between neurobiological and personal level explanation" [(3), p. 21, emphasis added].

Integrated explanations of the promised kind are said to be unavailable, in principle, to the isolationist FP view: this is precisely because to adopt the latter's "space of reasons" idea enforces an absolute distinction between the intentional and the mechanical.

To see what makes cognitive theory so appealing, it is worth getting clear about what exactly the FP view allegedly cannot do. Gerrans (3) accuses its proponents of operating with a disunified framework – one in which mechanisms are assumed to make only a causal difference to cognitive goings-on in a way that debars them from being properly explanatory (p. 15, 20). For example, Gerrans characterizes the FP view as being committed to the idea that organic damage might "*play a causal role* in introducing the drastic change in psychological structure but *plays no explanatory role*" [(3), p. 27, emphases added]. Does it make sense to think the explanatory space could carve up in the way Gerrans suggests? Can we distinguish between something's playing a merely causal versus a properly explanatory role? How should we understand this distinction?

As discussed in Section "Folk Psychology Rules," proponents of the FP view clearly allow that mental phenomena can be explained by what goes on in non-mental mechanisms. The FP view may be limited in that it is not interested in, or simply fails to provide, very detailed stories about the non-mental causal contributions of implementation mechanisms in information-processing terms. But it cannot be faulted for ruling out, or making it impossible to tell, such deeper explanatory stories. So this alone cannot be what makes its rival, the cognitivist view, special.[14]

Apparently, what makes cognitive theory special is that it brings something else – something quite unique – to the table. It regards the mind as a complex information-processing system – one that is organized in a hierarchical way, with a variety of interacting processes playing specific roles and where some of these diverse processes are responsible for the supervision of others in the system. Understanding the mind through the lens of cognitive theory allegedly provides a peculiar sort of intelligibility – one that allows theorists to go beyond the telling of merely "difference making" causal stories. The cognitive theory allows us to see how everything fits together in a systematic way; it bridges the gaps and enables explanations at many different scales and levels to be integrated by detailing how information flows from level to level and what role particular processes play in the wider cognitive economy [(3), p. 48, see also 32, 53, 79, 103].

From this vantage point, it is easy to see the attraction of having a broader vision of the mind that seeks to understand the roles played by various forms of cognitive activity, how various aspects of mind relate to and interact with one another, and how specific disturbances in those relations and interactions can lead to mental disorders with signature profiles.

This much is welcome. Yet friends of cognitivist SI, such as Gerrans (3), go further than this: they suggest that only cognitive neuroscience has what it takes to do the required integrating work. As Gerrans (3) says, "the *essential idea* of cognitive neuropsychiatry is that without a cognitive theory the problem identified by autonomy theorists … *cannot be*

[11] Gerrans (3) follows Murphy's lead of treating the SI view as best seen through the lens of a "minimalist cognitivism" (p. 18). Like Murphy, he sees that the future of mental health resides with brain sciences of the cognitive variety. He too regards psychiatry "as *a branch of cognitive neuroscience* by employing cognitive models that do not abstract away from, but are sensitive to, details of neural implementation" [(3), p. 37].

[12] On this vision, "personhood is a cognitive phenomenon constituted by the fact that personal-level phenomena, such as feelings, beliefs, emotions and desires arise at the highest levels of a cognitive processing hierarchy whose nature can be described and explained" [(3), p. 21]. Human cognition is thus "a complex hierarchy of computational processes performed by neural circuitry" [(3), p. 30].

[13] Motivating this proposal, with a Parthian shot at the perceived limits of the FP view, Gerrans (3) stresses that, "collecting and collating correlations between neural, phenomenological and cognitive properties of the delusional mind is useful but we need a theoretical approach that *fits all this information together*" (p. 14). It is here that we meet the idea that the tools of cognitive neuroscience are uniquely well suited to integrating "evidence from different disciplines about the way the mind configures itself in response to incoming information according to the way neural mechanisms influence cognitive processing" [(3), p. 14].

[14] There is great potential for confusion and conflation about just what the FP view and cognitive theory might, respectively, have to offer in terms of deeper explanations. As Section "Folk Psychology Rules" made clear, the FP view allows that we can go to a different level of description in order to get deeper explanations of mental disorders. Remarkably, in some places, Gerrans (3) talks in ways that suggest cognitive theory is wholly at peace with the FP view's suggestion that underlying neural mechanisms only ever explain by describing implementing mechanisms of cognitive phenomena. As he writes: "It is normal practice *to explain phenomena* such as amnesia or macular degeneration in terms of the way neural circuits *implement the cognitive processes* involved in memory or perception. This suggests that the way to explain psychology and phenomenology in terms of neurobiology is via a cognitive theory" [(3), p. 15, emphasis added]. If this were the whole story, it would be hard to distinguish what cognitive theory could offer that is really different from what the FP view offers. Yet, there are reasons to think this is neither the whole story about what cognitive theory has to offer nor the right one. In an unpublished paper, Murphy (forthcoming) upbraids Gerrans for talking about explanations of mental disorders by appeal to implementation mechanisms. By Murphy's lights, such talk is just an unfortunate hangover of the philosophical tendency to mix up analytic functionalism with cognitive psychology.

solved. The gap between neurobiology and psychology will be *unbridgeable*" [(3), p. 36]. Hence, "there *must be* an explanatory relationship between neuroscience and folk psychology" [(3), p. 33, emphasis added]. These are very strong, philosophically "musty" claims – and they are not self-evidently true.[15] We might well doubt that cognitive neuroscience *per se* is best placed to provide the desired integrating theoretical vision, especially in light of the concerns raised about Murphy's neurocentrism in the previous section.

What might persuade us that a brain-based cognitive theory is necessary to bridge the putative gaps? Allegedly, that cognitive neuroscience supplies special means for understanding the links between various mental phenomena. It can do so, again allegedly, precisely because it endorses a vision of the neurally housed mind "organised as a hierarchical system … which *uses representations* of the world and its own states to control behaviour" [(3), p. 47, emphasis added]. The claim is that cognitive theory posits neural representations that perform a variety of cognitive tasks and that once we understand how information flows between such representations, we will be in a position to provide complete and satisfying explanations of mental disorders in ways which make the links between subpersonal to personal-level cognitive phenomena intelligible. Thus, Gerrans (3) observes of this general strategy that, ultimately, supplying the correct account of what drives specific delusions requires accounting for "the way the brain encodes information acquired in experience and then reconstructs representations of that information when subsequently cued" (p. 33).

It seems that the central posits of cognitive theory – information and representation – provide the perfect theoretical glue for integrated explanations. Cognitive neuroscience promises to show how there can be relevant connections between various cognitive activities in a way that does not just cite correlations or brute causal relations. Instead, cognitive neuroscience alone proves to be genuinely explanatory of mental disorders because it alone *makes intelligible* multilevel interactions across various scales and levels.

Allegedly, cognitive neuroscience alone can achieve this feat because it is wedded to a representational theory of mind that assumes cognition to be at root both mechanical and intentional. Importantly, cognitive theory seems to provide us a new mark of the mental – not "rationality-in-intentionality" cast as person-level phenomena, to be sure. Instead it offers us a "content-in-intentionality" or CIT mark of the mental.[16] For those who accept something like CIT, even though the cognitive is regarded as quite a mixed bag that reaches across the so-called subpersonal and personal levels it is also united by the intelligible relations that are instantiated through the processing of informational and representational content in ways that define minds.

Some philosophers hold that cognitive scientists are committed to essentially characterizing minds in information-processing terms. This is, of course, not news. We frequently hear that

> cognitive science … has as its subject matter capacities like memory, perception, attention, language processing and reasoning. The concepts that cognitive sciences *take to be essential* for understanding their domain include information, representations, and algorithms [(19), p. 74, emphasis added].

Let us suppose, for the sake of argument, that Shapiro (19) is right in thinking that working cognitive scientists take themselves to use and need these kinds of conceptual tools. Would this help fans of the cognitivist SI view to justify the claim that cognitive neuroscience operates with unique explanatory tools that give it special gap-closing powers?

Would assuming CIT make cognitive neuroscience ideally well placed to provide gap-bridging solutions? One reason for thinking so is that CIT seems to imply the existence of something like a neurally based space of reasons. To posit a space of reasons mark II would be to assume that there exists a cognitive level at which various mental phenomena do not just brutally interact but intelligibly inter-relate because they communicate by trafficking in contentful information and representations. Content would, on this picture, be the shared common coin traded by all cognitive phenomena. The CIT picture seems to make it possible to understand cognitive relations in explanatorily illuminating ways that do not reduce to the giving of merely brutal, causal explanations.

Let us imagine that cognitive neuroscience posits a neural space of reasons, ala CIT, and embraces internalism about the vehicles of various mental contents. If so (assuming the above analysis is correct), it would follow that cognitive neuroscience would have utterly special resources for bridging the sort of gaps of which Gerrans (3) speaks. All that would have to be done to seal the deal would be to show in detail how the cognitive neuroscience, as imagined above, could use those resources to in fact close such gaps. Doing all of this would be an effective way of motivating an exclusively neurocentric version of the SI view.

Before assessing whether cognitive neuroscience, so construed, really has what it takes to close the said gaps, it is important

[15] For example, elsewhere, Gerrans speaks of the "*necessary* role of cognitive theory in linking the neurobiological and phenomenological levels of explanation" [(3), p. 18, emphasis added]. Methodologically speaking, it is strange that Gerrans (3) makes appeal to such general and wholesale philosophical justifications, for when pinning his philosophical colors to the mast he clearly tells us that: "*Murphy is right*. Our best understanding of the mind comes from understanding cognitive architecture. However that argument *cannot be established a priori for all mental phenomena*. The best we can do is construct, revise and, ultimately, unify case-by-case explanations" [(3), p. 14, emphases added]. This fits better with his more retail defenses of cognitive theory, such as when he claims that the "cognitive theory of visuo-motor control embedded in the overall architecture of cognitive control … is *required to explain* why high levels of activity in these regions produce loss of a sense of agency" [(3), p. 18, emphasis added], or when he tells us that "schizophrenic symptoms *can only be explained* in representational terms" [(3), p. 18].

[16] A CIT, intentionality-in-mechanism vision of the cognitive, is clearly in tune with the idea that, "The whole thrust of cognitive science is that there are sub-personal contents and sub-personal operations that are truly cognitive in the sense that these operations can be properly *explained only in terms of these contents*" [(18), p. 27, emphasis added].

to be clear about the source of the alleged need to do so. Notably, if there are any such explanatory gaps to bridge, then the need to bridge them is motivated by purely philosophical, not scientific, considerations. Without doubt, scientists and psychiatrists seek rich explanations of the roots of mental disorders. Providing such explanations would require going beyond FP and delving deeply into the sciences of the mind. However, crucially, providing such explanations is not the same as, nor does it require, bridging the putative intelligibility gaps – those that hold, e.g., between neurobiology and folk psychology and with which Gerrans (3) is concerned. Seen in this light, it becomes clear that Gerrans (3) seeks to motivate an exclusively cognitive neuroscientific take on SI by getting us to take seriously the need to address explanatory requirements of a distinctively philosophical kind.

The great irony is that elsewhere SIers reject the need to address such intelligibility demands as illegitimate. Compare the alleged need to make sense of the interactions between cognitive phenomena across levels by appeal to representational contents with the alleged need to make sense of the connections that hold between propositional attitudes in terms of rationality. If Gerrans (3) is right, cognitive theory can help us to make intelligible how various subpersonal cognitive phenomena inter-relate. How might it do this? By rendering the relations between cognitive phenomena intelligible. How? Not in RIT terms that explain how personal-level propositional attitudes relate rationally, to be sure, but in CIT terms that explain how neural representations relate contentfully.

It should now be easy to see why it would be a problem for SIers to advance this type of line. Any attempt to motivate a neurocentric cognitivist SI view by arguing that cognitive neuroscience alone can bridge otherwise unintelligible explanatory gaps requires being sensitive to the very sort of philosophical concerns that the SI view itself casts into doubt. Must there be some common feature (if not rationality then content) that is shared by all mental phenomena and which unifies them and explains how they intelligibly inter-relate? SIers say "No": They question the demand that "personal-level phenomena can only be explained in terms of other personal-level phenomena" [(3), p. 21]. This being the case, surely, we are also well within our rights to question whether there is a legitimate need for a unifying cognitive theory that makes intelligible how various cognitive phenomena intelligibly inter-relate in special, more-than-merely causal ways. As the old proverb reminds us, what's sauce for the goose is sauce for the gander.

And there is something else to consider. We might doubt that on close scrutiny appeals to information and representation could play the unifying and integrating roles that would satisfy the identified gap-bridging needs, if we were to take such needs seriously. The fact is that apart from bearing the names "cognitive," "representational," or "informational" nothing in so-called current cognitive theory deeply unifies all the various cognitive phenomena in terms of their importantly and interestingly diverse properties or roles.

Consider Gerran's claim that, "a scientist explaining some discrepant evidence *is doing the same thing* as the oculomotor system controlling the trajectory of a limb" [(3), pp. 46–7, emphasis added]. Is this credible? Undoubtedly, there may be some mileage in taking this route and drawing loose analogies for certain purposes. But a developed theory would be needed to back up any such claim if taken in a serious and literal way.[17]

To illustrate the point consider what Gerrans (3) has to say about the activation-information-mode (AIM) model of dreaming, which focuses on the flow of information within and between components of a control hierarchy. In discussing that model, he holds that the "*intrinsic cognitive properties* of these components *are preserved through transitions from mode to mode*. What changes are the interactions between these components" [(3), p. 79, emphases added].

Usually, in this sort of context, cognitive theorists upgrade talk of the flow of information to talk of the flow of informational content. Content, they hold, is what survives changes in mode and process. Canonically, the content of a mental state is determined by what it is about and how it represents the world to be. Now, if information processing literally involves the trading of contents that would make it easier to justify claiming that scientists and information-processing systems basically do the same cognitive work. Moreover, if content were the common coin that is always traded in some form, everywhere in the cognitive economy, it would be clear why there would have to be, and how there could be, intelligible relations holding between the many and various cognitive phenomena.

The trouble with this gambit is that it raises a host of unanswered questions. Just what is informational content anyway? What intrinsic cognitive properties does it have? Where does it get them? How can content be preserved through changes? How can it make a difference to cognition? How does it relate to representational content? Is it a kind of objective commodity? Does it make sense to say that we can take different perspectives – e.g., subjective and objective – toward it [as Ref. (3) appears to assume – see, e.g., p. 17]?

Cognitive theorists can avoid these tricky questions by sticking to an understanding of information in scientifically

[17] In the text surrounding this claim, Gerrans (3) makes clear that he is drawing on assumptions that predictive coding accounts of perception have made popular to support the idea that visual systems and scientists "do essentially the same thing." Predictive coding accounts understand cognition as a matter of making active inferences in continuous effort to minimize prediction error. But whether we should think of visual systems as really making contentful inferences at all, and whether if they do, they do so in anything like the way that scientists do, are highly contentious topics of current debate [for reasons why we ought to prefer a non-contentful reading of predictive processing, see Ref. (20, 21)]. For this reason, it might seem safer for cognitivists to advance a weaker claim about what makes these phenomena essentially the same. It might be argued that scientists and visual systems are essentially alike because they both use representations even though visual systems use different kinds of representation than scientists do. The idea here is that there is no requirement that visual systems and scientists need to operate with the same kinds of content in order to qualify as representational systems. While it is technically correct to go this way it raises afresh the question of what unifies and intelligibly relates these two cognitive phenomena if not the fact that they both involve manipulations of content of the same kind. The point is that without full details, it is far from clear why we should accept that vision and scientific theorizing – which appear to be quite disparate cognitive activities – are essentially alike.

respectable terms – those of covariance and correspondence. That is perfectly fine, but then it is difficult to justify claims that basic information processing is content involving in a way, which would license drawing a strong analogy with the theoretical activity of scientists.

This is just one example of a disunity objection to the integrationist picture. To make a full dress case against such a vision would require a much longer discussion [(10), esp. ch. 4]. For our purposes, it suffices to note that anyone offering a bridge building, unifying cognitive theory must answer the sorts of questions raised above. *Prima facie*, it seems they will only be able to do so with the backing of a well-developed naturalistic theory of content.

To highlight why such a theory is needed, consider a different set of cases. In many of the explanations that Gerrans (3) offers of delusions the "felt" aspect of the phenomenon in question turns out to be a pivotal factor. The phenomenological and emotionally charged aspects of our experience apparently matter to and help explain some of the strong tendencies we have when responding to, interpreting, and accounting for our situations. Yet, as is notoriously well known, we currently lack anything like a workable theory that shows how we are to understand such qualitative phenomena in purely information processing or representational terms. Once again, it looks like disunity rather than unity is the word of the day.

Things are even more puzzling if we consider the roles imaginings are meant to play in the integrative explanations on offer by cognitive theorists, such as Gerrans (3). For example, he holds that simulative activity generates imaginings that can be incorporated in a wider cognitive economy. By this, he means that imaginings can be the basis for action (including mental action). Despite the fact that imaginings are influential and we often act on them, they are cognitively interesting and distinct because they lack many of the properties of canonical propositional attitudes, such as belief [(3), p. 18].

On this score Gerrans (3) tells us that

> Imagination uses the mind's cognitive resources, such as perceptual, doxastic and emotional processing to create simulations. It thus inherits the intentional structure of these counterpart processes. However *qua* simulations *imaginative states do not have congruence conditions*. [(3), p. 105].

The basic claim, which is plausible enough is that imagination deploys specialized neural circuitry to "construct and manipulate *representations which have representational contents but no congruence conditions*" [(3), p. 114, emphasis added]. Gerrans (3) is concerned to show that simulative imagining can figure in and make a difference to one's thinking without the content of such imaginings being believed.

Yet, since most theorists hold that mental content requires some kind of correctness or congruence condition, it is puzzling in what sense imaginations can be said to have representational content if they lack such conditions altogether in the way Gerrans (3) proposes. What remains if you subtract congruence conditions from a mental representation? Gerrans's (3) answer is intentional structure. But it is not clear what exactly puts the intentionality in this structure for cognitivists if not the existence of mental representations with congruence conditions.[18]

Let us be clear. A simulative account of imaginings is attractive for many reasons [(23), ch. 4]. However, it is far from clear that imaginings without congruence conditions are best understood as any kind of mental representation for precisely the reasons stated above (20, 22). But even if this proves possible it would remain unclear how a simulative account of the imagination that emphasized the lack of congruence conditions could contribute to a unified cognitive theory of minds.

Our capacity for producing narratives – often quite spectacular ones – is yet another place in which it is important to recognize that interesting forms of cognition have special properties that break the standard representationalist mold. Gerrans (3) proposes that particular forms of delusional thinking arise from signature breakdowns in the usual interactions between cognitive systems. These breakdowns in turn prompt patterns of default thinking that take the form of experientially charged imaginative episodes. Default thoughts of this stripe provide raw material that can be woven together into what are, for those in the grip of a delusion, spectacular and hypersalient narratives. Importantly, such default thoughts "are subjectively adequate responses to experience constructed as narrative elements or fragments" [(3), p. 101].

The basic idea is that when operating in the default mode, we assemble first pass, coherent stories. Yet even when these stories are internally coherent, they are not always subjected to critical epistemic scrutiny. According to Gerrans (3) when unsupervised by decontextualized systems, the products of default thinking are not scrutinized for consistency or veridicality; they are not evaluated against "competing narratives for accuracy or utility" [(3), p. 77].

This is hardly surprising since the great bulk of narratives do not aim at truth. Although narratives all share certain basic structural properties, we must look to the contexts in which we use a given narrative in order to determine its semantic properties. Thus, as Goldie (24) points out "Fictional narratives do not aspire to be true, whereas real life narratives do. A narrative is fictional not in virtue of its content being false, but in virtue of its being narrated, and read or heard, *as part of a practice of a special sort*" (pp. 152–3, emphasis added). Thus fictional narratives, offered up as fictions, invite "the audience to imagine or make believe that what is being narrated actually happened, even when it is known that it did not. Thus the question of reference and of truth simply does not arise within the 'fictive stance'" [(24), pp. 152–3]. For these reasons, Goldie concludes that, "reference and truth have no application in fiction, but do have application in historical and everyday explanation" [(24), p. 154]. Different kinds of narratives exhibit different kinds of semantic

[18]When thinking about what might be leftover in such a subtraction, it is useful to consider Gerrans's (3) claim that, "different cognitive processes have *different computational properties* that enable them to meet their congruence condition. These *properties provide the intentional structure of representations* produced by different cognitive processes. For example, the representations produced by the visual system are 3D coloured scenes derived by processing spectral and luminance information" [(3), p. 105, emphases added]. However, once no correctness conditions are in play, it is not clear in what sense the residual structures ought to be thought to bear representational content or even what it means to say that they do (22).

properties, and we understand these differences if we are alert to the roles that these different kinds of narratives play in our lives and thinking.

A crucial contrast becomes evident if we compare the uncritical use of narratives that do not aim at truth with the intense critical scrutiny of beliefs and claims that do. In the most serious cases, the latter are subject to the norms of scientific testing, where we seek to fix what we believe only "according to standards of consistency and empirical adequacy" [(3), p. 13]. Put simply, some forms of cognition do have representational contents and do play roles in our cognitive economy that make them subject to epistemic norms which simply do not apply to other forms of cognition. Other forms of cognition lack these features. Our so-called default thoughts – those generated when our minds are wandering or in screensaver mode are a prime example. They do not involve any "attempt to confirm an empirical hypothesis" [(3), p. 76].

Although the above analysis only scratches the surface, the important thing to note is that the detailed explanations Gerrans (3) offers of the complexities of delusional thinking gain their power by focusing on the way diverse cognitive phenomena (e.g., feeling, imagining, and narrating) interact in virtue of their special cognitive roles and properties. Contrariwise, these explanations gain nothing from making the additional cognitivist assumption that all mental phenomena are united because they are, somehow, representational in character.

To tell a convincing explanatory story about minds and how they can become disturbed in particular ways, we need to recognize the important diversity of mental phenomena rather than insisting on a cognitivist account that downplays those differences in favor of fulfilling a philosophically motivated demand for unity. We can relinquish CIT and its problematic intelligibility requirement, recasting the integrating cognitive theory in far less ideologically demanding ways than do the friends of cognitivist SI. This does not mean we should give up on understanding how various mental phenomena interact or that we should not seek to understand the roles they play in the larger cognitive economy. It simply means that we can make sense of the relevant interactions and relations between mental and other phenomena without insisting that informational and representational content are needed to account for the intelligibility of such relations.

Only if we fully free ourselves from the constraints of FP-based philosophical suppositions about what is necessary for something to count as a properly cognitive phenomena does it become possible to concoct accounts of cognition that are truly unconstrained by FP thinking about the basic nature of minds. Interestingly, radically enactivist approaches that lay stress on the importance of interactions over contentful representations as the common coin of the cognitive looks well placed to pick up the explanatory burden (10). This is especially so if it is accepted that "what needs to be explained here is not just the causal interactions among neurons but *the way those interactions* enable cognitive processes and experiences" [(3), p. 30].[19]

KEEPING FP IN THE PICTURE

Only once the siren songs of an exclusively brain-focused future vision for psychiatry are silenced can the ground for a suitably open-minded and philosophically uncontaminated rendering of the SI view be laid. This closing section shows that when modestly formulated in the way suggested above, the SI view can make peace with an unimperialistic vision of FP.

Consider, once again, Gerrans's (3) plausible suggestion that narrative-based and theory-based explanations differ in important ways because they answer to different epistemic standards. Thus, to understand the delusional mind requires understanding how these modes of cognition interact or fail to interact in particular conditions.

Let us assume that Gerrans's (3) answer is along the right lines. Let us assume, for example, that those under the sway of specific delusions do indeed construct stories as opposed to rationally evaluated beliefs in order to make sense of such episodes. We might wonder, assuming they are not natural-born narrators, how they come to be able to weave such stories? We might be interested to know why a given kind or genre of story rather than another is more compelling to some populations rather than another? Or, why – upon experiencing an underlying mismatch between what-is-felt and what-was- anticipated-would have-and-should-have-been-felt – the narratives of deluded people unfold in one standard variant rather than another. The thing to notice is that in order to explain and understand key features of delusional narratives and the narrative practices that enable their generation requires looking at socioculturally and not purely neural factors [for extended arguments along these lines, see (23, 25–27)]. This is especially the case when it comes to understanding the distinctive kinds of norms relevant to the sorts of cognitive activity that differentiate narrative from scientific practices.

The point is that understanding the relevant norms requires looking beyond the brain (27). Only outside the skull of individuals do we find what we need for making sense of normative features of the cognitive phenomena that need explaining. Yet it is also when we look to certain public practices that we come by the resources for adopting a softer take on FP and its role in therapy. Certain kinds of treatment urge us to make best use of tools already available within cultures – such as incorporating traditional narrative practices into therapy – in order to respond to those in need.

There are compelling reasons to agree with Gerrans (3) that our foremost ways of making sense of ourselves and others are grounded in explanations that are not theoretical but narratively based. Such explanations function, primarily, as normalizing explanations. In giving them, any of a number of explanatorily relevant factors might be cited (e.g., facets of X's character, X's mood, X's larger projects, the content of this or that propositional attitude of X, and so on). Crucially, like historical explanations, these folk psychological explanations are not general and abstract but take the form of narratives that emphasize details that are personal and particular.

[19] Enactivists, of course, encourage multi-stranded investigations, involving explanations that are pitched at various "levels" and "scales." Gerrans (3) acknowledges this. Taking the case of vision as a prime example, he emphasizes the need for theories that seek to simultaneously investigate different levels of cognitive activity and how they integrate. Yet here he notes, "Even enactive theorists of vision who disagree with Marrians nonetheless debate with them about the causal relevance of mechanisms at different levels" [(3), p. 43].

Consider an idealized test case. Imagine a person suffering from a psychiatric condition that is brought on by a cascade of factors rooted in neural causes. Imagine that the condition can be wholly and successfully addressed by a perfectly targeted neuroscientific intervention. Even in this imagined case – one that best favors a purely neurocentric vision of psychiatry in terms of diagnosis explanation, and treatment – it is plausible there would be a need for the person to achieve a rehabilitating self-understanding. Graham (1) captures this point when he says:

> to mend or heal from a disorder in a self-respecting and dignified manner requires discovering a positive or purposeful place for past and present episodes of disorder in the … course of a person's life … [this] often consists of dealing with conflicting interpretations of one's past … [(1), p. 14].

The take-home lesson is that even in ideal cases in which targeted neural inventions might wholly relieve specific conditions we should not typically expect psychiatric therapy to boil down to a simple business of eradicating "disease" in the way a narrowly construed medical model can suggest.

Murphy (2) appears prepared to acknowledge that there is a need for psychiatry to go beyond the brain, at least in some cases. In this vein, he states clearly that, "there are important roles for non-scientific thinking about the methods of psychiatry" [(2), p. 47].[20]

Importantly, even for those who press for a more thorough brain-based vision of psychiatry, there need be no conflict between endorsing both an SI view of the field and recognizing the need and importance of non-scientifically focused therapies.[21] When it comes to therapy, a cautious pluralism seems to be the appropriate stance: it appears we need a variety of approaches if we are to improve the situation of individuals. Individual therapeutic requirements need to be assessed on a case-by-case basis. What matters is that a pluralist approach is always possible – and typically desirable – when it comes to treatment. As Murphy (2) rightly stresses, "Even if we have established that a symptom is best explained in terms of one main causal factor, such as neurotransmitter abnormality, it does not follow that treatment must be directed at directly manipulating that causal factor" [(2), p. 369].

This is all well and good, but Murphy (2) is almost completely silent about which non-scientific approaches and forms to therapy might be usefully brought to bear. And it is here that folk psychological narrative practices are likely to play a central role. This is because narratives are the familiar, everyday medium through which most of us readily evaluate and reflect upon our reasons, attitudes, and situations (24). Reviewing and recasting our narratives, with the assistance of others, is not only a way of making sense of our lives in new and fresh ways it can open up possibilities for living them differently[22]. Narrative practices afford such new possibilities precisely because they provide a means for thinking afresh about "who we are" based on richer understandings of our peculiar situations by revisiting our possible pasts and reimagining our possible futures.

Understanding FP as a kind of narrative practice in this way connects perfectly with the ambitions of narratively based therapies that seek to use so-called "talking cures" to empower people in the construction of a viable "future trajectory rather than achieving past accuracy" [(1), p. 14]. FP, as a special kind of narrative practice, is a possible object of philosophical and scientific study in a way that is wholly compatible with the modest rendering of the SI view argued for in the previous section. The views are compatible because FP, construed as a narrative practice, is not to be understood as a general theory embedded in that practice from which a philosophically discernable mark of the mental that defined mental disorders is to be sourced.[23]

CONCLUSION

There are excellent reasons to resist a forced choice between standard format FP and SI views of psychiatry. On the one hand, in its original variant, the FP view attempts to provide a definitive mark of all that is properly mental, which is imperialistic and isolationist. On the other hand, the SI view, at least when formulated in its popular cognitivist version, is unjustifiably and potentially unhelpfully overly narrow and neurocentric. Consequently, adopting either of these views of psychiatry in their standard forms threatens to leave us with an ideological vision of psychiatry's future that is too extreme and too limited. A better way forward is to salvage what is best from heavily edited versions of the familiar versions of the FP and SI views on the market, combining what remains to best effect.

Ultimately, the arguments presented here have been pitched at a quite general level, whereas to make good on this plan for reconciliation in a wholly convincing manner requires more detailed philosophical work on case studies in ways, which it

[20] Crucially, however, Murphy (2) insists that it is important to "distinguish between supplementing and replacing the medical model" (p. 367).

[21] Narrative therapy is, for example, neither scientifically focused nor scientifically based. It uses special techniques in order to provide the tools for empowering people, enabling them to exercise their agency in wider and more positive ways. Narrative therapy, although very much in the mould of "talking cures," is thus unlike more familiar psychoanalytic approaches to therapy in that it does not seek to divine and understand past causes of current trauma. Nevertheless, there seems no reason to discount narrative therapy as a bone fide therapy given that it is has been used successfully to help people deal better with a wide range of psychiatric and traumatic conditions, including asthma, anorexia, bulimia, and depression.

[22] Hutto DD, Gallagher S. Re-authoring narrative therapy: opening the way for future developments. *Philos Psychiatr Psychol* (Forthcoming).

[23] Whether FP should be understood as a narrative practice as opposed to, and distinct from, a theory of mind remains a controversial matter of dispute in the literature. It would take too much space to attempt to settle the issue or detail all the consequences of going one way rather than the other, in this paper. Extended arguments for treating FP as a narrative practice that does not reduce to theory can be found in Ref. (23, 25–28, Hutto DD, McGivern P. Updating the story of mental time travel: narrating and engaging with our possible pasts and futures. In: Altshuler R, Sigrist MJ, editors. *Time and the Philosophy of Action*. London: Routledge (Forthcoming)).

has not been possible to provide in this paper. But if the above arguments are sound, then the ambitions of this paper will have been realized and the ground will have been laid for those future, follow-up endeavors.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This work was supported by the Australian Research Council Discovery Project, Embodied Virtues and Expertise (DP: 1095109); the Marie-Curie Initial Training Network, TESIS: Towards an Embodied Science of InterSubjectivity (FP7-PEOPLE-2010-ITN, 264828); and the (Ministerio de Economía e innovación) Spanish DepartÂment of Economy and Innovation: Agency, Normativity and Identity: the Presence of the Subject in Actions (FFI-2011-25131).

REFERENCES

1. Graham G. *The Disordered Mind: An Introduction to Philosophy of Mind and Mental Illness*. London: Routledge (2009).
2. Murphy D. *Psychiatry in the Scientific Image*. Cambridge, MA: MIT Press (2006).
3. Gerrans P. *The Measure of Madness*. Cambridge, MA: MIT Press (2014).
4. Davidson D. Problems in the explanation of Action. In: Smart JJC, Pettit P, Sylvan R, Norman J, editors. *Metaphysics and Morality: Essays in Honour*. Oxford: Blackwell (1987). p. 35–49.
5. Davidson D. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press (1984).
6. Lewis DK. How to define theoretical terms. *J Philos* (1970) **67**:427–46. doi:10.2307/2023861
7. Lewis DK. Psychophysical and theoretical identifications. *Australas J Philos* (1972) **50**:249–58. doi:10.1080/00048407212341301
8. Jackson F. *From Metaphysics to Ethics*. Oxford: Oxford University Press (1998).
9. Carruthers P. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press (2011).
10. Hutto DD, Myin E. *Radicalizing Enactivism*. Cambridge, MA: MIT Press (2013).
11. Stich S. Autonomous psychology and the belief-desire thesis. *Monist* (1978) **61**(4):699–718. doi:10.5840/monist197861446
12. Fodor JA. Methodological solipsism considered as a research strategy in cognitive science. *Behav Brain Sci* (1980) **3**:63–73. doi:10.1017/S0140525X00001771
13. Hohwy J. *The Predictive Mind*. Oxford: Oxford University Press (2013).
14. Hohwy J. The self-evidencing brain. *Noûs* (2014). doi:10.1111/nous.12062
15. Rowlands M. Arguing about representation. *Synthese* (2015). doi:10.1007/s11229-014-0646-4
16. Hutto DD, Kirchhoff M, Myin E. Extensive enactivism: why keep it all in? *Front Hum Neurosci* (2014) **8**:706. doi:10.3389/fnhum.2014.00706
17. Murphy D, Smart G. Review of the disordered mind: an introduction to philosophy of mind and mental illness, George Graham. *Notre Dame Philosophical Reviews*. Routledge (2010). Available from: https://ndpr.nd.edu/news/24392-the-disordered-mind-an-introduction-to-philosophy-of-mind-and-mental-illness/
18. Seager W. *Theories of Consciousness*. London: Routledge (1999).
19. Shapiro L. When is cognition embodied? In: Kriegel U, editor. *Current Controversies in Philosophy of Mind*. London: Routledge (2014). p. 73–90.
20. Hutto DD. REC: revolution effected by clarification. *Topoi* (2015). doi:10.1007/s11245-015-9358-8
21. Orlandi N. *The Innocent Eye: Why Vision Is Not a Cognitive Process*. Oxford: Oxford University Press (2014).
22. Hutto DD. Overly enactive imagination? Radically re-imagining imagining. *South J Philos* (2015) **53**(S1):68–89. doi:10.1111/sjp.12122
23. Hutto DD. *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press (2008).
24. Goldie P. *The Mess Inside: Narrative, Emotion and the Mind*. Oxford: Oxford University Press (2012).
25. Hutto DD. Folk psychology as narrative practice. *J Conscious Stud* (2009) **16**(6–8):9–39.
26. Hutto DD. ToM rules, but it is not ok. In: Costall A, Leudar I, editors. *Against Theory of Mind*. Basingstoke: Palgrave (2009). p. 221–38.
27. Hutto DD, Kirchhoff MD. Looking beyond the brain: social neuroscience meets narrative practice. *Cogn Syst Res* (2015) **34–35**:5–17. doi:10.1016/j.cogsys.2015.07.001
28. Hutto DD. Narrative understanding. In: Carroll N, Gibson J, editors. *The Routledge Companion to Philosophy of Literature*. London: Routledge (2016). p. 291–301.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

HYPOTHESIS AND THEORY
published: 07 December 2015
doi: 10.3389/fpsyt.2015.00171

Against Explanatory Minimalism in Psychiatry

*Tim Thornton**

College of Health and Wellbeing, University of Central Lancashire, Preston, UK

The idea that psychiatry contains, in principle, a series of levels of explanation has been criticized not only as empirically false but also, by Campbell, as unintelligible because it presupposes a discredited pre-Humean view of causation. Campbell's criticism is based on an interventionist-inspired denial that mechanisms and rational connections underpin physical and mental causation, respectively, and hence underpin levels of explanation. These claims echo some superficially similar remarks in Wittgenstein's *Zettel*. But attention to the context of Wittgenstein's remarks suggests a reason to reject explanatory minimalism in psychiatry and reinstate a Wittgensteinian notion of levels of explanation. Only in a context broader than the one provided by interventionism is that the ascription of propositional attitudes, even in the puzzling case of delusions, justified. Such a view, informed by Wittgenstein, can reconcile the idea that the ascription mental phenomena presupposes a particular level of explanation with the rejection of an *a priori* claim about its connection to a neurological level of explanation.

Keywords: Campbell, levels of explanation, intentionality, mechanism, rationality, Wittgenstein

Edited by:
Derek Strijbos,
Radboud University, Netherlands

Reviewed by:
Daniel Douglas Hutto,
University of Wollongong, Australia
Dawa Ometto,
Utrecht University, Netherlands

***Correspondence:**
Tim Thornton
tthornton1@uclan.ac.uk

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry

Received: *29 July 2015*
Accepted: *23 November 2015*
Published: *07 December 2015*

Citation:
Thornton T (2015) Against Explanatory Minimalism in Psychiatry. Front. Psychiatry 6:171. doi: 10.3389/fpsyt.2015.00171

INTRODUCTION

Psychiatry deals with phenomena that range between large-scale higher order social phenomena (e.g., poverty, cultural norms), person level phenomena (e.g., trauma, symptoms such as delusions and syndromes such as depression), and the sub-personal level phenomena (e.g., genes, neurones). It advances explanations that invoke a variety of factors across the scale and both proximal and distal. It is tempting to think that this apparent heterogeneity could be simplified, in principle at least, by fitting it into a picture of different levels of explanation – whether ontological or epistemological – which relate together in some general ways.

The actual applicability of this picture to present psychiatry has been contested (1). Typically, psychiatry trades across levels. This paper, however, describes a principled attack on the very idea of levels of explanation in favor of a form of explanatory minimalism. Put roughly, causal explanation can trade across putative levels because causation is brute and answers to no *a priori* conditions of intelligibility. That being so, the very idea of a "level of explanation" – which depends on such *a priori* assumptions about intelligibility – is undermined. Or so John Campbell argues in a number of papers (2–4).

Having set out Campbell's argument for explanatory minimalism for psychiatry, I compare it to some similar sounding remarks from Wittgenstein's collection *Zettel* (5). Like Campbell, Wittgenstein denies the necessity for mechanisms to mediate (apparently) causal connections and also denies the assumption that rational connections between mental phenomena need be mediated by underlying neurological mechanisms. But Wittgenstein's remarks are aimed at undermining mechanistic accounts of the intentional directedness of mental states, not at denying that there

is a characteristic level of explanation for mental phenomena. The problem lies not with the idea of levels of explanation but with an unwarranted metaphysical assumption about how they relate.

The final section builds on Wittgenstein's account of the normative connections between mental phenomena and argues that Campbell's explanatory minimalism is insufficient for psychiatry because it provides no account of what constitutes states as mental states, which plays an important role in psychiatric explanation.

BACKGROUND: LEVELS OF EXPLANATION

There are two dominant approaches to the idea of levels of explanation: ontological and epistemological. Both attempt to shed light on the idea of levels of explanation by characterizing the differences between the levels and also the constraining relations between them.

The ontological view is part of a traditional reductionist picture of the world. On this picture, sciences of the mind, such as psychiatry and psychology, can in principle be reduced to biology (which might be construed as physiology or evolutionary biology), biology to chemistry, and chemistry to physics. Oppenheim and Putnam expressed this view in their classic 1958 paper "Unity of science as working hypothesis."

> It is not absurd to suppose that psychological laws may eventually be explained in terms of the behaviour of individual neurons in the brain; that the behaviour of individual cells – including neurons – may eventually be explained in terms of their biochemical constitution; and that the behaviour of molecules – including the macromolecules that make up living cells – may eventually be explained in terms of atomic physics. If this is achieved, then psychological laws will have, *in principle*, been reduced to laws of atomic physics… [Ref. (6): p. 407]

Oppenheim and Putnam go on to argue that the unity of science is served by "microreductions." These are reductions in which:

> The objects in the universe of discourse of [the reduced science or theory] are wholes which possess a decomposition into proper parts all of which belong to the universe of discourse of [the reducing science or theory] [ibid: 407].

In fact, they argue more strongly that microreduction is the *only* method seriously available for the unity of science [ibid: 408]. They then go on to explore the consequences of this view by examining the preconditions for successfully attaining unity via microreduction. Since microreduction is construed as the only serious possibility for the unity of science, and since its success rests on a number of other things being the case, the goal of unification has a number of presuppositions:

> 1. There must be several levels.
> 2. The number of levels must be finite.
> 3. There must be a unique lowest level …
> 4. Anything of any level except the lowest must possess a decomposition into things belonging to the next lowest level … [ibid: 409].

This list suggests the following view of nature and the constraining relations between levels of explanation. The world is made up of basic building blocks or atoms, which display regularities that can be described in the law statements of the most basic science. The basic atoms also combine to form larger structures that display characteristic regularities of their own. These can in turn be codified in the law statements of higher level sciences. But the higher level regularities do not emerge out of nothing. They can be explained as the consequences of the more basic patterns of behavior of atoms. So, the structure of the world and the structure of science can be seen as two isomorphic hierarchies of levels.

The picture suggests three interrelated mutually reinforcing views of the levels of explanation. First, they correspond to different disciplines within science. Second, higher levels contain objects that are constituted from lower level objects. Third, higher level objects are larger than lower level objects. These views fit together on the assumptions that different sciences study objects at different scales and that objects only interact with other objects at the same level. However, these assumptions have been criticized (7).

There is, however, another and quite different approach to levels of explanation, which has been influential. It is based not on the size and composition but rather degree of abstraction to higher order causal processes. This is Marr's threefold epistemological distinction between:

> *Computational theory*: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
> *Representation and algorithm*: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
> *Hardware Implementation*: How can the representation and algorithm be realized physically? [Ref. (8): p. 25]

This hierarchy does not concern different ontological levels but rather different ways of understanding the same ontology. The highest, most abstract level concerns the function of a system. It might be carried out by a variety of different algorithms at the middle level. Finally, the same algorithm might be realized in different physical ways at the lowest level. Thus, the higher levels are multiply realizable by lower levels. Determining the computational level is a matter of determining the goals of a system independently of its physical or neurological properties.

Although Marr's epistemic approach seems appropriate for its original application to vision, where the goals of the system can be theorized about independent of algorithm and physiological realization, it is less clear that it applies to psychiatry. As Dominic Murphy argues, actual practice in psychiatry is

to determine the functions of systems in part with a view of what the lower level physiology could sustain. "[O]ur understanding of realisation feeds back into and constrains our understanding of the abstract demands of cognition" [Ref. (1): p. 105].

Neither, however, does the ontological view of levels of explanation fit psychiatry because "causes described in genetic vocabulary will be related to effects described in terms of behavioural tests, for example, and generalisations will cross levels" [ibid: 108]. Thus, according to Murphy, there are different systems operating at different levels, unlike the epistemic view, but the different levels interact, unlike the ontological view.

Murphy's argument starts from a relaxed approach to the nature of levels – "I have little to say about what levels of explanation actually are" [ibid: 103] – and then argues that they do not apply to psychiatry. Whatever they are, psychiatric explanation typically crosses them. Such an argument, however, leaves open the response that it merely reflects the current imperfect state of psychiatry. It is tempting to think that Oppenheim and Putnam's picture reflects how reality must be structured even if, for contingent reasons, causal generalizations can link different levels. Equally, the fact that knowledge of physical realization informs more abstract theories of function need not conflict with the idea that there are, in principle, different levels of abstraction applicable to a completed psychiatry. These possibilities remain because Murphy's arguments do not, explicitly at least, undermine the intelligibility of the concept of levels of explanation. By contrast, according to John Campbell, the very idea of a level of explanation is a reflection of a mistaken pre-Humean view of causation. There is less to explanation than the requirement to fit a specific level would require.

CAMPBELL'S CRITICISM OF LEVELS OF EXPLANATION IN PSYCHIATRY

To characterize his target, Campbell gives the example of a discussion of thought insertion by Christopher Frith (9). Frith claims that whether or not inappropriate firings of dopamine neurons are found in subjects who experience thought insertion, this fact could not be used to explain their experiences. It would shed no light on why that kind of symptom, rather than another, was produced by inappropriate firings of dopamine neurons. To shed light, Frith assumes that we need an account pitched at a particular level: in Frith's case that of a sub-personal but still cognitive model of mechanisms supposedly responsible for thought insertion.

Campbell suggests that the assumption that there is a right level of explanation that clarifies things in the way Frith desires is the result of a pre-Humean view of causal explanation. Although often forgotten, Hume successfully argued that there need be no intelligible connection between cause and effect. That is implicit in his rejection of any logical connection to analyze the apparently necessitating relation between cause and effect. Causal connections are merely brute facts to be discovered by experience.

Resisting the idea that the right kind of cause and effect have to be intelligibly, rather than merely brutely related also undercuts the motivation for the levels of explanation picture on both approaches: ontological and epistemic.

> We naturally seek a certain kind of intelligibility in nature; we naturally try to find explanations that will show the world to conform to reason, to behave as it ought. Hume's point is that there are no such intelligible connections to be found. This point has generally been accepted by philosophers thinking about causation. Hume's comments nonetheless do leave us in an uncomfortable position, because we do tend to look for explanations that make the phenomena intelligible to reason. We are prone to relapse, to think that after all we must be able to find intelligibility in the world. This tendency survives, I suspect, in the idea of 'levels of explanation'. The idea is that within certain levels of explanation, we will find a particular kind of intelligibility. [T]he lesson from Hume is that there is no more to causation than arbitrary connections between independent variables of cause and effect. We have to resist the demand for intelligibility [Ref. (2): p. 201].

This is not just a restatement of Murphy's claim that, in psychiatry, explanations may cross levels. Rather, the very idea of levels of explanation, understood as causation operating under some constraint of intelligibility, is itself undercut. This applies to both ontological and epistemic versions as both assume that causation is governed by *a priori* constraints, whether degree of abstraction or composition.

This leaves, however, the issue of shedding some light on the nature of causal connections (if not *a priori* light on particular causal connections). In (non-mental) cases of causation, the notion of *mechanism* plays a central role in empirical research. Searching out the way in which causal influence is transmitted has been an important part of scientific practice. "It would seem a kind of madness if someone were to acknowledge that there is a causal link, but propose that there may be no mechanism linking the two" [Ref. (3): p. 138]. But if science has usefully explored the mechanisms that mediate causal influence, there must be some paradigmatic mechanisms that stand in need of no further explanation and the transmission of motion by impulse, in Hume's billiard ball example, is one such prototype.

Nevertheless, the idea that there *must* be such a mechanism is a kind of synthetic *a priori* claim which, Campbell suggests, should be rejected in line with Hume's argument. He adopts the interventionist model defended most extensively by James Woodward in *Making Things Happen* according to which for X to be a cause of Y is for intervening on X to be away of intervening on Y (10). The rejection of the necessity of a mechanism and the adoption of an interventionist approach opens up the possibility of a causal connection – in accord with interventionism – where there is no mechanism. In the case of psychiatry, however, the key issue is causation in the absence of a *mental* mechanism, whatever that is taken to be.

Just as we find it natural to expect there to be a mechanism underpinning material causal connections – even if this assumption lacks any genuine *a priori* justification – so Campbell also suggests that in the case of mental causation we expect there to

be a rational connection between propositional attitudes. The rational link between two propositional attitudes is our paradigm of a mental causal mechanism. So, if one hears someone explain that they believe that Tranmere Rovers won their most recent football match because they heard it on the BBC, which they take to be trustworthy, no further inquiry is needed as to why the beliefs about what they hear and trust cause the belief about the result. Again, however, while the idea that mental causation is underpinned by rational connections is natural and compelling, it lacks *a priori* justification.

> [T]here is an analogy between:
> 1 the idea that propositional attitude ascriptions depend on the ascription of rationality to the subject, and
> 2 the idea that all causal interactions between pieces of matter must be comprehensible in mechanistic terms.
>
> Both ideas express an insight – that we find it extremely puzzling when we encounter causal relations among propositional attitudes that are not broadly rational, just as we find it extremely puzzling when we encounter causal interactions between physical objects that are not mechanistic, and that involve spooky 'action-at-a-distance'. Both ideas express a natural impulse of philosophers – to elevate this kind of point into a kind of synthetic *a priori* demand that reason makes on the world. This impulse has to be resisted [Ref. (3): p. 142].

In both cases, there is a genuine insight. As a matter of custom and habit, we find an absence of material mechanisms and an absence of rational connections between mental states puzzling. But in both cases, it is a characteristic philosophical error to promote this natural expectation into a justified *a priori* claim that the world must respect. Mere custom and habit cannot rationally sustain any such demand on how the world must be.

The rejection of the necessity for rational connections between causally related mental states looks to ease a central problem for the philosophy of psychiatry: explaining delusions. There need be nothing genuinely mysterious about a causal connection, which lacks a rational connection (the expected mental mechanism).

> Suppose you believe:
> 1 that this man is stroking his chin, and
> 2 that this man believes you need to shave.
>
> What is it for the first belief to be a cause of the second? On the interventionist analysis, it is for the intervention on the first belief to be a way of changing whether you have the second belief. So if some external force changed your belief that this man is stroking his chin, you would no longer believe that he believes you need to shave. There is no appeal to rationality here, no appeals to mechanism [Ref. (3): p. 143].

The causal connection between one state and another is underpinned in interventionist terms based on the idea that if intervening on the first belief is a stable way of bringing about a change in the second then this is sufficient for there to be a causal connection between them.

Spelling this idea out involves a little more complexity, however. Given a scanner capable of yielding a complete microphysical description of the human body and a longitudinal study of schizophrenia in a population, Campbell suggests that it might be possible to form a disjunctive characterization of the set of microphysical states that are nomically sufficient for schizophrenia. But that function from physical states to illness would lack any concise expression and would not be couched in terms of variables, which could be affected by local intervention. This point reflects the pragmatic aspect to interventionism: not every nomically sufficient state counts as a cause.

For propositional attitudes to count as causes of delusions, Campbell suggests two conditions have to be met. There should be "systematic relations between cause variables and the subsequent delusion" and there should be a correlation between a change of the cause and a change of the effect [Ref. (3): p. 146]. More generally for the causal explanation of mental states, the causal variables, which he calls "control variables," should have large, specific, and systematic correlations with their effects akin to the way the controls of a car systematically control its behavior. These conditions do not require a rational connection, however. To repeat Campbell's phrase, there need be "no appeal to rationality here."

> The classical philosophical approach has been to regard propositional attitudes as part of a 'conceptual scheme' that we bring to bear in describing the ordinary world. This conceptual scheme is taken to have strong *a priori* constraints on its applicability. In particular, as we have seen, rationality is taken to be a norm with which the scheme has to comply. The appeal I have just been making to the notion of a control variable is intended to replace this invocation of rationality. [I]t is the fact that we have control variables, not the fact that we have rationality, which means that we are 'at the right level' to talk of beliefs and desires [Ref. (3): p. 147].

The phrase "at the right level" occurs in inverted commas to flag the fact that the notion of the right explanatory level has been undercut. Without a pre-Humean insistence on the intelligibility of causal relations, there is no more to the notion of being at the right level than that there is a causal relation tracked through the idea of control variables.

With the idea of control variables replacing an *a priori* requirement for rationality in mental causation, psychiatric explanation of delusions is in principle in the same predicament as the explanation of any other belief. Causal explanation has been achieved once one has an understanding of the variables necessary for changing the delusional belief entertained. The apparently principled problem of attempting to fit primary delusions into some sort of rational framework is replaced by a practical problem of charting the variables that affect them. But is that minimal approach enough for psychiatric explanation?

WITTGENSTEIN ON CAUSATION AND MECHANISM

Campbell suggests a mutually supportive analogy between the denial that mental causation requires rational mediation and that physical causation requires a mechanism. The latter denial echoes some remarks by Wittgenstein in *Zettel*. In this section, I will outline the context of Wittgenstein's discussion, outline a key disanalogy and hence begin to suggest a reason to reject explanatory minimalism in psychiatry.

The later Wittgenstein makes a number of comments both explicitly and implicitly on the connection between mind and body. Throughout his various discussions of propositional attitudes, he denies the possibility of an explanation of meaning or forming an intentional mental state via an appeal to brain states. This accords with his criticisms of causal and dispositional explanations of rule following in the *Philosophical Investigations* (11). As the discussions of both real and ideal machines imply, the attempt to explain rules by appeal to mechanisms is either question begging or fails to sustain their normativity [ibid §§193–4]. Thus, no account could be given in which thought processes might be read off from brain processes.

Such considerations might be thought to motivate the following claim in *Zettel*:

> No supposition seems to me more natural than that there is no process in the brain correlated with associating or with thinking; so that it would be impossible to read off thought-processes from brain-processes. I mean this: if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the *system* continue further in the direction of the centre? Why should this order not proceed, so to speak, out of chaos? [Ref. (5): §608]
>
> It is thus perfectly possible that certain psychological phenomena *cannot* be investigated physiologically, because physiologically nothing corresponds to them [ibid: §609].

These passages could be interpreted as merely denying that the systematicity of thought can be explained as resulting from an underlying systematicity in the brain. In other words, they could be interpreted as a denial of reductionist explanations of meaning and mental content.

This interpretation would also be consistent with another passage:

> Imagine the following phenomenon. If I want someone to take note of a text that I recite to him, so that he can repeat it to me later, I have to give him paper and pencil; while I am speaking he makes lines, marks, on the paper; if he has to reproduce the text later he follows those marks with his eyes and recites the text. But I assume that what he has jotted down is not *writing*, it is not connected by rules with the words of the text; yet without those jottings he is unable to reproduce the text; and if anything in it is altered, if part of it is destroyed, he sticks in his 'reading' or recites the text uncertainly or carelessly, or cannot find the words at all. – This *can* be imagined! – What I called jottings would not be a *rendering* of the text, not so to speak a translation with another symbolism. The text would not be *stored up* in the jottings. And why should it be stored up in our nervous system? [ibid: §612]

This passage does not say that the marks on paper do not form a system. It is just that they do not form a system of the same sort as writing. That is why they are not a *rendering* of the text. They are not connected by *rules* to words. But in that case, what is their connection to the text supposed to be? Given that this is supposed to be an analogy for the connection between the nervous system and our linguistic abilities, one suggestion is that the marks are connected to written or spoken words *causally* rather than via shared meaning. If this were the case, while the internal system could not be used to *reduce* mental content, it could still play a necessary causal role.

But in fact, Wittgenstein goes further than this. He suggests that there need be *no* cause of a memory in the nervous system. Nothing need be stored "up there" in any form. There need be no physiological regularity or order causing psychological order. Mental order could proceed out of chaos:

> The case would be like the following – certain kinds of plants multiply by seed, so that a seed always produces a plant of the same kind as that from which it was produced – but *nothing* in the seed corresponds to the plant which comes from it; so that it is impossible to infer the properties or structure of the plant from those of the seed that it comes out of – this can only be done from the *history* of the seed. So an organism might come into being even out of something quite amorphous, as it were causelessly; and there is no reason why this should not really hold for our thoughts, and hence for our talking and writing [ibid: §608].
>
> I saw this man years ago: now I have seen him again, I recognise him, I remember his name. And why does there have to be a cause of this remembering in my nervous system? Why must something or other, whatever it may be, be stored up there *in any form*? Why *must* a trace have been left behind? Why should there not be a psychological regularity to which *no* physiological regularity corresponds? If this upsets our concepts of causality then it is high time they were upset [ibid: §610].

These two passages are pitched against a model of levels of explanation in which a psychological regularity corresponds to a regularity at a lower level. That is, they run counter to the assumption in both Putnam and Oppenheim's ontological, but also Marr's epistemological, accounts of the constraints operating between levels of explanation. By contrast with Campbell, Wittgenstein does not reject the idea that there is a characteristic level of explanation for mental happenings.

Using terminology not in widespread use when Wittgenstein wrote these remarks, they amount to the claim that plants' development do not supervene on their seeds' microstructure. Given contemporary understanding of RNA, this might seem a bizarre possibility. But the denial is of a *modal* claim: that the plant *must* be determined by something in the seed's structure. Wittgenstein denies this assumption, natural though it currently seems.

Like Campbell, Wittgenstein denies that the psychological regularity has to be mediated by one at the neurological or physical level. Furthermore, like Campbell, he suggests an analogical denial. What we might have taken to be a causal physical connection – in the seed and tree example – also need not be mediated by any mechanism. (The analogy is with psychological regularity depending on the physical level.) There is a difference between Wittgenstein and Campbell, however, in that Wittgenstein here assumes the very connection between causation and mechanism that Campbell denies in favor of interventionism. Wittgenstein asks:

> Why should there not be a natural law connecting a starting and a finishing state of a system, but not covering the intermediary state? (Only one must not think of *causal efficacy*.) [ibid: §613]

This passage assumes that the denial of an intervening mechanism implies a denial of causal efficacy. Given that the natural law would sustain the kind of intervention conditionals, then a particular kind of seed producing a particular kind of plant would count as a causal connection according to Campbell. But this looks merely like a difference of terminology. Both Campbell and Wittgenstein can grant a law-like connection. Campbell's account suggests it should count as "causal" while Wittgenstein denies it "causal efficacy." Both deny that there need be an intermediate mechanism. But aside from its causal status, the metaphysical facts are agreed.

Despite that, however, there is a fundamental difference. Wittgenstein's remarks are aimed at removing the tension of reconciling a connection made at the mental level in mental and, according to him, non-causal terms with assumptions about underlying causal mechanisms at a physiological level. Campbell's, by contrast, suggest that transitions at the mental level can, when explanatory, be causal. I will now explore the significance of this difference.

THE ROLE OF AN APPEAL TO RATIONALITY IN WITTGENSTEIN'S DISCUSSION OF INTENTIONALITY

In order to see the difference between Campbell and Wittgenstein, it will be helpful to start with what they share. There are two particularly clear examples in the *Philosophical Investigations* where Wittgenstein, like Campbell, rejects an appeal to underlying mechanisms to explain a connection at the mental level. But, by contrast with Campbell, he goes on to suggest a different account of the mental connection. It is this that suggests Wittgenstein's commitment to a characteristically mental level of explanation.

One example concerns the ability to read out loud, of what reading comprises. He considers the temptation to identify the ability with a mechanism through the example of a comparison between an expert reader and a beginner who can only read words by laboriously spelling them out.

> Now we would, of course, like to say: What goes on in the practised reader and in the beginner when they utter the word *can't* be the same. And if there is no difference in what they are currently conscious of, there must be one in the unconscious workings of their minds, or, again, in the brain. – So we'd like to say: There are, at any rate, two different mechanisms here! And what goes on in them must distinguish reading from not reading. – But these mechanisms are only hypotheses, models to explain, to sum up, what you observe [Ref. (11): p. §156].

Rejecting the hypothetical mechanism – whether an unconscious mental mechanism or physiological one – as well as conscious experiences of being guided or feelings of familiarity, he stresses instead the relation between the text and spoken words, however, mediated. Whatever mediating processes there may be are not what is meant by "reading."

A second example concerns the intentional directedness of having someone in mind.

> "I am thinking of N." "I am speaking of N."
> How do I speak *of* him? I say, for instance, "I must go and see N. today" – But surely that is not enough! After all, when I say "N.", I might mean various people of this name. – "Then there must surely be a further link between my words and N., for otherwise I would *still* not have meant him." Certainly such a link exists. Only not as you imagine it: namely, by means of a mental *mechanism* [ibid: §689].

In the surrounding discussion, various putative explanatory connections are considered and rejected including the idea that no such connection exists, that it is created in being verbally avowed (and that both are true!), and that it is connected to what would, counter-factually, have been reported. Wittgenstein's discussion fits a meta-philosophical injunction: "The point is not to explain a language-game by means of our experiences, but to take account of a language-game" [ibid: §655]. But it also accords with a brief assertion in the middle of an earlier discussion of the intentional directedness of propositional attitudes: "It is in language that an expectation and its fulfilment make contact" [ibid: §445].

This terse comment picks up the idea that avowals and descriptions of expectations and other propositional attitudes reuse the same fragments of language as descriptions of the events that would satisfy them (12). To be able to form such a propositional attitude requires the contingent ability to fit one's avowals and actions into the rational pattern articulated in language. The criticism of underlying mechanisms is made against the background account that psychological order has a rational linguistically mediated structure.

This suggests a fundamental contrast with Campbell's view. Although both Campbell and Wittgenstein reject mechanisms, Wittgenstein's rejection goes hand in hand with a normative and rationalistic view at the mental level which Campbell, at least in the series of papers so far discussed, downplays. In the final section, I will outline the consequences of this disagreement for causal explanation in psychiatry. But first I will briefly summarize how Wittgenstein's views of meaning and mental content suggest a picture of levels of explanation.

A WITTGENSTEINIAN VIEW OF LEVELS OF EXPLANATION

I began by outlining the two dominant approaches to thinking about levels of explanation, both ontological and epistemic. Both approaches not only suggest ways of distinguishing levels but both also suggest constraining relations of either composition, in the ontological case, or realization, in the epistemological approach. Wittgenstein's discussion of mental phenomena in, especially, his *Philosophical Investigations* sets out some of the key differences between normative meaning-related or intentional connections and causal connections. But his remarks in *Zettel* run counter to the assumptions, particularly in Putnam and Oppenheim, of the constraining relations between the psychological and the neurological.

In other words, Wittgenstein's remarks suggest a middle ground between Campbell, on the one hand, and Putnam and Oppenheim, on the other hand. Thinking that there are distinct forms of intelligibility need not imply an *a priori* view of a constraining relation between them. Putnam and Oppenheim assume a series of levels of explanation but then impose a reductionist view of their relations. Campbell rejects the intelligibility of levels of explanation in the first place. Wittgenstein, however, suggests that grasping events or states as mental phenomena presupposes fitting them into a normative and rational linguistic structure but denies that this necessitates connections to a non-normative pattern of causal relations. This suggests that to understand a state to be a state of expectation, for example, involves relating it in a characteristic way to events that would satisfy or fulfill it and hence to presuppose a particular *a priori* pattern of intelligibility. But Wittgenstein denies the need, *a priori* at least, to connect this to any underlying pattern of neurological cause and effect.

The denial of an *a priori* connection to underlying neurology is not the same as denying an *a posteriori* connection. The remarks in *Zettel* do not contradict the possibility of neurological and psychiatric research establishing local connections between medical interventions and psychological effects. Instead, they caution merely against assuming that a pattern at one level must be relatable to a pattern at a lower level.

INFLATING EXPLANATORY MINIMALISM IN PSYCHIATRY

Earlier I reported Campbell's claim that:

> [I]t is the fact that we have control variables, not the fact that we have rationality, which means that we are 'at the right level' to talk of beliefs and desires [Ref. (3): p. 147].

I asked whether the resulting picture of explanatory minimalism was sufficient for psychiatric explanation. It is not sufficient because it provides no account of what constitutes a state as a belief, or a desire or even a delusion. In the absence of that, however, psychiatric explanation would miss a key feature of the phenomena it aims to illuminate.

In an earlier paper, Campbell himself endorses the role of rationality as a presupposition for holding propositional attitudes. He suggests two general reasons for this. The less important one is as follows.

> One simple reason for thinking that rationality is critical here is that unless you assume the other person is rational, it does not seem possible to say what the significance is of ascribing any particular propositional state to the subject. If you tell me that someone rational thinks that it is raining, then given that the person is rational and does not want to get wet, I know what kinds of behavior to expect. If, however, the person is not at all rational, then saying they have the belief has no implications at all for how they will behave [Ref. (13): p. 89].

Campbell's focus is on the *ascription* of propositional attitudes to others. The imputation of rationality goes hand-in-hand with an ascription of propositional attitudes. The argument in the passage seems to concern what follows from the ascription. Without the assumption that the subject is also rational, it is not clear what can be inferred from the ascription of particular mental states to them. But this argument surely broadens. Without a rational pattern, the very idea that the subject has some determinate mental state is undermined (14–16).

There is a second connection, however, which Campbell thinks is the more important. It concerns the connection between rationality, belief, and meaning. Understanding others' utterances and hence ascribing beliefs to them is only possible against a background assumption of rationality. There is a balance between possible irrationality and the ascription of meaning.

> The finding of irrationality can always be traded for a finding of mistranslation. And we should always translate so as to find the subject rational in the use of a term by the lights of the subject's own understanding of the term [ibid: 90].

This sketch of the connection between interpretation and the ascription of belief echoes Wittgenstein's suggestion of a linguistic mediation of mental states and their intentional objects. The very idea of having propositional attitudes presupposes a harmony between the meaning of utterances, the mental states held, and the pattern of actions they rationalize.

Thus, the claim that control variables, rather than rationality, constitutes the "right level" to talk of beliefs and desires fails to address a prior constitutive question. What is it about some particular causes and effects, described using the interventionist model of causation, which constitutes them as intentional mental states in the first place? Given its broad application to causation in the non-mental as well as mental

world, talk of control variables alone is insufficient to address this question. But introducing issues of language, interpretation, and rationality suggests that there is a particular level of explanation which is of central importance to psychiatry when it addresses the meaning and content of psychological phenomena.

In the example described above, Campbell argues that for the belief that this man is stroking his chin to cause the belief that this man believes you need to shave all that is needed is a suitable interventionist counterfactual relation rather than an appeal to rationality. But without some further background conditions, of which rationality is one plausible candidate, the ascription of determinate mental states is illicit.

It may seem, however, that defending the role of a rational connection between utterance, mental state, and behavior is particularly difficult in the case of psychiatric explanation. After all, psychiatry investigates phenomena that appear to resist rational understanding. While this is true and puzzling, however, it does not threaten the connection itself.

Consider Campbell's discussion of Capgras in the earlier paper (13). He uses the link between meaning and rationality to suggest a problem with the interpretation of characteristic expressions of the delusion. The characteristic type of utterance associated with the delusion is: "That woman is not my wife!." But that sentence might be used to make a number of different claims. It might, for example, be used to flag the discovery of illegality in a past wedding ceremony. The most plausible interpretation in the context of the expression of the Capgras delusion is something like: This [demonstrated] woman is not that [remembered] woman. But such an interpretation is put under strain because, typically, the subject of the delusion does not attempt to carry out any of the paradigmatic or canonical forms of checking appropriate for such a claim: for example, discussing past events and checking memories. They do not do what they ought to do to check such a thought. Given the link between meaning, mental content, and rationality, this apparent failure of rationality undermines such an interpretation.

Campbell himself goes on to try a partial accommodation of the delusion within rational space by suggesting it might be a deviant hinge or framework proposition since, if it were, it would be rational not to subject it to testing. It is unclear whether this approach can work as it is unclear what understanding there can be of a framework proposition which is not shared (17). But the difficulties Campbell highlights seem genuine. Does a subject who makes a paradigmatic Capgras utterance but does nothing else different really believe that their partner is an imposter? Likewise, does the Cotard utterance "I am dead" really express the impossible belief that the subject is dead? The difficulty seems fundamental to such cases.

In a more recent paper, Campbell seems more pessimistic about fitting delusions into any sort of rational pattern. He considers a delusion in which the subject thinks that her mother's thoughts were inserted into her mind via raindrops and the air conditioner. He points out that the structure of this delusion could not be used to teach what is meant by "rationality." But further:

> The trouble is not even that the patient is not rational. We have no idea what a rational way of going on would be, once one has accepted that thoughts are being inserted into one's mind. How must the world be, for that to happen? Would it make sense to argue with this patient that, by her own lights, it is not the raindrops in the air conditioning that should be blamed, but rather the electrical sockets all around? We have departed so far from the ordinary world that we have no idea what stands fast and what has to go [Ref. (3): p. 141].

Again, these seem to be genuine and substantial difficulties in working out what the subject actually thinks. But Campbell offers a particular interpretation of the difficulty. He says:

> We should not appeal to the idea that there are *a priori* constraints on causal relations among propositional attitudes. We have to accept that the propositional attitudes are one thing and the causal relations among them are another. If propositional attitudes do not conform to rationality, that is puzzling. But we cannot legislate in advance that this cannot happen [ibid: 140].

This seems an unjustified response, however. The problem is not merely that there are contingent breakdowns in the expected rational connections between identifiable propositional attitudes. Rather, in the case of delusion, the nature of the supposed propositional attitudes themselves is, and continues to be, puzzling. Hence, for example, attempts to suggest that the delusion may be a propositional attitude of imagination rather than belief [e.g., Ref. (18)]. It is not that the bizarre quality of delusions threatens the general connection between meaning, mental state, and rationality but instead that the general connection helps to illuminate what is so puzzling about delusion. The connection to rationality is not arbitrary: it helps justify the claim that a state is a mental state or that an utterance expresses a particular propositional attitude.

CONCLUSION

Given the heterogeneity of the factors that feature in explanations in psychiatry, it is tempting to assume that, in principle, they can be related within an ordered hierarchy of levels of explanation. There is reason, however, to doubt that this picture fits contemporary psychiatry. But that leaves open the response that that is a reflection merely on the current state of psychiatric research and that a completed psychiatry would form an ordered hierarchy.

More radically, John Campbell has argued in recent papers that the very idea of levels of explanation presupposes a discredited pre-Humean view of causation. He claims that although the assumption that physical causation is mediated by mechanisms and that psychological causation is mediated by rational relations have both been fruitful neither need to be true. With their rejection as synthetic *a priori* claims about the world, the idea of levels of explanation also falls away to leave an explanatory minimalism.

Comparing Campbell's remarks with some superficially similar remarks in Wittgenstein's *Zettel* suggests an objection to explanatory minimalism. The very idea of a state being a mental state presupposes broader connections. Rationality is one such candidate. If so, explanation in psychiatry inflates from Campbell's minimalism and introduces an appropriate level of explanation at which mentality comes into view. But it is possible to hold on to the necessity of such general levels of explanation while rejecting *a priori* claims about how different levels of explanation must relate to each other.

REFERENCES

1. Murphy D. Levels of explanation in psychiatry. In: Kendler KS, Parnas J, editors. *Philosophical Issues in Psychiatry*. Baltimore, MD: Johns Hopkins University Press (2008). p. 102–25.
2. Campbell J. Causation in psychiatry. In: Kendler KS, Parnas J, editors. *Philosophical Issues in Psychiatry*. Baltimore, MD: Johns Hopkins University Press (2008). p. 199–216.
3. Campbell J. What does rationality have to do with psychological causation? Propositional attitudes as mechanisms and as control variables. In: Bortolotti L, Broome M, editors. *Psychiatry as Cognitive Neuroscience*. Oxford: Oxford University Press (2009). p. 137–50.
4. Campbell J. Causation and mechanisms in psychiatry. In: Fulford KWM, Davies M, Gipps R, Graham G, Sadler J, Stanghellini G, et al., editors. *Oxford Handbook of Philosophy and Psychiatry*. Oxford: Oxford University Press (2013). p. 935–49.
5. Wittgenstein L. *Zettel*. Oxford: Blackwell (1981).
6. Oppenheim P, Putnam H. Unity of science as a working hypothesis. In: Boyd R, Gasper P, Trout JD, editors. *Philosophy of Science*. London: MIT Press (1991). p. 405–27.
7. Craver C. *Explaining the Brain*. Oxford: Oxford University Press (2007).
8. Marr D. *Vision*. San Francisco, CA: W.H.Freeman (1982).
9. Frith C. *The Cognitive Neuropsychology of Schizophrenia*. Hove: Lawrence Erlbaum (1992).
10. Woodward. *Making Things Happen*. Oxford: Oxford University Press (2003).
11. Wittgenstein L. *Philosophical Investigations*. Oxford: Blackwell (1953).
12. Arrington RL. Making contact in language: the harmony between thought and reality. In: Arrington RL, Glock H-J, editors. *Wittgenstein's Philosophical Investigations*. London: Routledge (1991). p. 175–202.
13. Campbell J. Rationality, meaning, and the analysis of delusion. *Philos Psychiatr Psychol* (2001) **8**:89–100. doi:10.1353/ppp.2001.0004
14. Davidson D. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press (1984).
15. Dennett D. *The Intentional Stance*. Cambridge, MA: MIT Press (1987).
16. Hopkins J. Wittgenstein, Davidson, and radical interpretation. In: Hahn F, editor. *The Library of Living Philosophers: Donald Davidson*. Chicago: Open Court Hopkins (1999). p. 255–85.
17. Thornton T. Why the idea of framework propositions cannot contribute to an understanding of delusion. *Phenomenol Cogn Sci* (2008) **7**:159–75. doi:10.1007/s11097-007-9079-6
18. Currie G. Imagination, delusion and hallucinations. *Mind Lang* (2000) **15**:168–83. doi:10.1111/1468-0017.00128

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

HYPOTHESIS AND THEORY
published: 18 April 2016
doi: 10.3389/fpsyt.2016.00057

What Is Constructionism in Psychiatry? From Social Causes to Psychiatric Classification

Raphael van Riel[1,2]*

[1] Philosophy, Philipps-University Marburg, Marburg, Germany, [2] Philosophy, University Duisburg-Essen, Essen, Germany

It is common to note that social environment and cultural formation shape mental disorders. The details of this claim are, however, not well understood. The paper takes a look at the claim that culture has an impact on psychiatry from the perspective of metaphysics and the philosophy of science. Its aim is to offer, in a general fashion, partial explications of some significant versions of the thesis that culture and social environment shape mental disorders and to highlight some of the consequences social constructionism about psychiatry has for psychiatric explanation. In particular, it will be argued that the alleged dependence of facts about particular mental disorders and about the second order property of being a mental disorder on social facts amounts to a robust form of constructivism, whereas the view that clinician–patient interaction is influenced by cultural facts is perfectly compatible with an anti-constructivist stance.

Keywords: philosophy of science, explanation in psychology, metaphysics, social construction, psychiatric classification

OPEN ACCESS

Edited by:
Leon De Bruin,
VU University Amsterdam,
Netherlands

Reviewed by:
Anna Welpinghus,
Technical University Dortmund,
Germany
Lieke Asma,
VU University Amsterdam,
Netherlands

***Correspondence:**
Raphael van Riel
raphael.vanriel@uni-due.de

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry

Received: *07 October 2015*
Accepted: *24 March 2016*
Published: *18 April 2016*

Citation:
van Riel R (2016) What Is Constructionism in Psychiatry? From Social Causes to Psychiatric Classification. Front. Psychiatry 7:57. doi: 10.3389/fpsyt.2016.00057

INTRODUCTION

It is common to note that social environment and cultural formation shape mental disorders. For instance, the Surgeon General David Satcher states in the preface to *Mental Health: Culture, Race, and Ethnicity* (Supplement) that

> [t]he cultures from which people hail affect all aspects of mental health and illness, including the types of stresses they confront, whether they seek help, what types of help they seek, what symptoms and concerns they bring to clinical attention, and what types of coping styles and social supports they possess. Likewise, the cultures of clinicians and service systems influence the nature of mental health services. [(1): preface]

The details of the claim that social or cultural facts or events have a significant or systematic impact on mental disorders are, however, not well understood. Some construe it as a purely epistemological claim (2), on other occasions it is mentioned without any further explication. For instance, in the supplement *Mental Health: Culture, Race, and Ethnicity* just quoted, it is stated that

> [c]ultural differences in the expression and reporting of distress are well established among American Indians and Alaska Natives. These often compromise the ability of assessment tools to capture the key signs and symptoms of mental illness [...] Words such as "depressed" and "anxious" are absent from some American Indian and Alaska Native languages [...]. Other research has demonstrated that certain DSM diagnoses, such as major depressive disorder, do not correspond directly to the categories of illness recognized by some American Indians. Thus, evaluating the need for mental health care among American Indians and Alaska Natives requires careful clinical inquiry that attends closely to culture. [(1): chapter 4.]

We will turn to similar examples below. Typically, such claims are regarded as articulating versions of *social constructivism*, as opposed to objectivism; authors quickly move from the idea that sociocultural environment has an impact on mental disorders to constructivist rhetoric. The paper takes a look at the claim that culture has an impact on psychiatry from the perspective of metaphysics and the philosophy of science. Its aim is to offer, in a general fashion, partial explications of versions of social constructivism about mental disorders and to show which claims regarding a sociocultural influence on mental disorders amount to social constructivism, and which claims do not. In particular, I will discuss constructivist claims about mental disorders themselves, such as posttraumatic stress disorder (PTSD), about the subjective experience, the phenomenology, and symptoms that are indicative of a disorder, about the second order property of being a mental disorder, instantiated by, for instance, PTSD or autism, and about constructivism about aspects of the clinician–patient interaction. It will turn out that social constructivist rhetoric, in many cases, does not amount to social constructivism, properly construed.[1]

Social constructivism in psychiatry can be tentatively characterized by contrasting it with what one may want to call *radical objectivism* (a view I will use for illustrative purposes only).[2] According to radical objectivism about psychiatry, types of mental disorders, and the type *mental disorder* itself, are just like the types the natural sciences deal with, in that they are in some sense *explanatorily independent* of social facts or events.[3] Neither do facts about mental disorders have to be explained in terms of underlying social facts nor should the occurrence of mental disorders be explained by reference to social causes, in a sense to be specified.[4] This is what the dispute between objectivism and constructivism is about. Compare events of evaporation of water and what they, as such, depend on with events of elections and what elections, as such, depend on. Neither do explanations of events of *evaporation of water* in physics or chemistry cite social causes nor is the evaporation of water, in these sciences, explained in terms of underlying social facts. Although the evaporation of water in the ocean may depend on social facts, namely, the social causes of global warming, in an explanation of what evaporation of water *is*, we should not cite the social causes of actual evaporation of water. By contrast, it seems reasonable to assume that *elections* are to be explained in terms of social causes, such as joint decisions to vote, or decisions to hold an election by individuals or groups who have a certain social status within a society. Moreover, unlike facts about the evaporation of water, the fact that the election takes place, facts about how it develops, etc., explanatorily depend on other social facts – facts about actions of individuals that count, in the relevant context, as votings. Social constructivism about psychiatry assumes that facts about mental disorders are, with respect to what they, as such, explanatorily depend on, a bit-like facts about elections. Radical objectivism assumes that facts about mental disorders are, in this respect, more like facts about the evaporation of water. They may sometimes be caused by social facts, but this is irrelevant when it comes to understanding what they are.

Social constructivism is widespread; in some circles, it may even be regarded as trivially true. So, why bother? The aim of this paper is not to *defend* or *argue against* social constructivism; it aims at a clarification of what social constructivism about mental disorder consists in, or *may* consist in. It will turn out that in the literature, one can find different versions of social constructivism. As these forms of social constructivism are often only implicit, part of the work will consist in uncovering some hidden constructivist commitments.

The paper proceeds as follows. The first section introduces some basic elements of social constructivism, including an elaboration on the difference between causal and non-causal, or, as I will also sometimes say, *metaphysical* explanation or explanatory dependence.[5] The section "Versions of Social Constructivism about Psychiatry" sketches, in an abstract way, the various versions of social constructivism about psychiatry, and introduces the different targets of constructivist claims, such as mental disorders, their symptoms, and the property of being a mental disorder. Each of the remaining Sections ["Social Constructivism about Mental Disorders", "Social Context, Experience, Phenomenology, and Symptoms", and "Mental Disorder and Social Norms"] deals with a particular version of social constructivism about psychiatry. Some consequences for psychiatric explanation and, thus, for psychiatry as a science are highlighted in the conclusion. Note that the conclusions drawn in this paper are somewhat preliminary, in three respects. First, the way I will present the different explications of social constructivism is non-committal as to the metaphysical details of social constructivism. From the perspective of metaphysics, this may appear dissatisfying. But in order to pave the way for a more thorough theory of social constructivism in psychiatry, we should remain neutral on some of the metaphysical details. Second, the conclusions drawn below are preliminary in that the distinctions discussed here may not exhaust the field. I have focused on versions of social constructivism that appear to surface within prominent areas in philosophy and psychiatry. And although the taxonomy offered here is inspired by systematic

[1] Throughout this paper, I will ignore the possible cultural impact on experiments in psychiatric research, for two reasons. First, I doubt that it requires special treatment, if constructivism about experiments in psychiatry poses a problem at all. If social environment significantly shapes experimentation in psychiatry, it will do so in other sciences as well, and for similar reasons. By contrast, the targets of social constructivism discussed in this paper are, prima facie, *special*. Second, I am not aware of any form of constructivist rhetoric about experimentation in psychiatry that is relevantly distinct from constructivist rhetoric that shows in descriptions of patient–clinician interaction.

[2] Social facts, whatever these are, are no less *objective*, in the sense of "*real*," than non-social ones, on the view endorsed here. It is a distinction inside naturalism. Social facts are part of nature in that they depend on natural facts about individual brains. This is an articulation of my conviction – most of the points made in the paper are compatible with more liberal views.

[3] I use the term "types," in line with the literature, to designate what is signified by predicates, not to designate predicates or concepts expressed by predicates.

[4] Some such view can be ascribed to Kendell (3) and Boorse (4), at least about the second order property of being a mental disorder, a similar view is defended by Kendler et al. (5), who admit that social aspects may have an impact on disorders, but who assume that mental disorders can be individuated in a way similar to types in the natural sciences.

[5] There is a vast literature on the metaphysics of these forms of dependence; for introductory texts, see Ref. (6); for recent work on metaphysical dependence and explanation, see, for instance, the papers published in Ref. (7).

considerations about candidate versions of social constructivism, I will not offer an argument to the effect that the versions of social constructivism discussed below exhaust the field. Finally, each of the versions discussed below deserves further attention. The paper offers partial explications of versions of social constructivism. Some aspects will be left out.

Basic Tenets of Social Constructivism

Social constructivism about psychiatry is opposed to radical objectivism. Both claims concern the subject matter of psychiatry, the types of objects and the connections psychiatry deals with. The present section introduces the conceptual tools of social constructivism.

A quick look at social constructivism in other areas of philosophy will offer a more thorough idea of what social constructivism about some subject matter consists in. A prominent example that has been extensively discussed in the past years (8–11) is social constructivism about institutions or institutional facts. Social constructivism about institutions holds that institutions (or institutional facts) depend on specific intentional states of individuals. For instance, on this view, the fact that some sea-shell, dollar-bill, or coin is money depends on the fact that people collectively accept it as money [famously argued by Searle (8, 9)]. This form of social constructivism is characterized by its *target* (facts about social institutions, such as money), the relevant facts upon which its target is supposed to depend, or the target's *social grounds* (in this case: facts about collective acceptance), and the *relation* that is supposed to hold between the two – in this case: some form of *metaphysical dependence*.

Here is the second example. Feminist philosophers have suggested that in some sense, gender is socially constructed (12, 13). On one version of this thesis, the idea is that women and men *become women* and *men* (in at least one significant sense of these terms) not due to their hard-wired biological make-up, but rather due to social causes, such as among other things, shared expectations on the side of caregivers, discursive practices (repeatedly marking the distinction between boys and girls/men and women), and esthetic practices within a society. Again, this form of social constructivism can be characterized in terms of its target (the occurrence of gender identities in individuals), the alleged grounds of this target (causal influences, such as expectations and discursive practices), and the connection between the two – here, *causation*.

Versions of social constructivism about psychiatry can be characterized in a similar format – in terms of their *target*, the alleged *grounds* of the target (I use "ground" for both, causal and non-causal grounds), and the *relation* that is supposed to hold between the two. We need to distinguish between two types of dependence relations – causal and non-causal dependence. As the reader might not be familiar with this distinction, let me illustrate the difference by way of some examples. An avalanche that occurs due to an earthquake is *caused* by the latter – and the avalanche can be causally explained by reference to the earthquake. When a rock hits a window so that the window shatters, the fact that the window shatters can be causally explained by reference to the fact that to rock hit the window. And when the heating of water leads to a transition from liquid to vapor, then the vaporization of water can be causally explained in terms of the heating of water. Causal explanations involve a temporal component – causes precede their effects. By contrast, the objects involved in *non-causal* explanations do not necessarily stand in a temporal successor relation. The existence of a forest depends on the existence of trees, and that there is a forest can be explained by reference to the presence of trees. But the existence of the trees does not cause the forest to exist, and the existence of the trees need not precede the existence of the forest. The hole in the Swiss cheese depends on the cheese; and the existence of the hole can be explained by reference to features of the cheese. But it need not be the case that there was, first, the cheese and then the hole. On naturalistic accounts of the mind, the processing of visual information depends on particular physiological processes. These processes underlie the processing of visual information, and the latter can be explained in terms of the former. But the physiological processes do not cause the processing of visual information and need not precede it.

"Because"-statements typically express explanations. Some explanations are causal, others are not. When speaking of dependence in what follows I mean *explanatory* dependence. An intuitive understanding of the difference between causal and non-causal explanation along these lines, in terms of examples and based on the observation that causal explanation essentially involves a temporal component non-causal explanation does not require, is sufficient for our present purposes. In the present context, non-causal dependence will sometimes be referred to as "constitution."

As already indicated, versions of social constructivism may, in general, differ with respect to the relation they postulate between the social grounds of their targets and these targets. Versions of social constructivism fall into two categories – those that credit social facts with a relevant causal role for the etiology of the given target and those according to which the target non-causally depends on social facts. These categories mirror, to some extent, the following distinction drawn by Sally Haslanger:

> *Causal Construction*: Something is causally constructed iff social factors play a causal role in bringing it into existence or, to some substantial extent, in its being the way it is.
>
> *Constitutive Construction*: Something is constitutively constructed iff in defining it we must make reference to social factors (13, p. 98).[6]

This distinction applies in the context of theorizing about psychiatry. Consider the claim that some mental disorders depend on social norms (we will turn back to this view below, Section

[6]It is worth noting that the claim about causal construction should be interpreted as a claim about types, rather than tokens. To return to our example from the introduction: the actual evaporation of water in the ocean is caused by social factors, and the way it proceeds also hinges on these social causes; but evaporation as such is independent of social factors; its being actually caused by social factors is irrelevant when it comes to understanding what it consists in, namely, some form of phase state transition. By contrast, to understand what gender categories consists in, social factors that played a role in the causal development of gender identities are in fact relevant. Why this is so is a question that transcends the boundaries of the present paper; it is a general question regarding classification in the social sciences.

"Mental Disorder and Social Norms") – that reference to social norms is relevant in an explanation for why a patient suffers from a mental disorder. Social norms may have a causal impact on the development of disorders, and facts about social norms may ground, in the non-causal sense, facts about disorders.

To illustrate, consider Stier's [(2), p. 28] interpretation of Wakefield's critique of current diagnostic practices in the case of anxiety disorders. Wakefield explains the fact that "current criteria allow diagnosis when someone is, say, intensely anxious about public speaking in front of strangers" by reference to "American society's high need for people who can engage in occupations that require communicating to large groups" [(14), p. 154]. Stier appears to suggest that considerations like these reveal that there is a "[normative] impact of society on the concept of mental disorder" [(2), 28f.]. Although I doubt that Wakefield's considerations concerning anxiety disorders support any such view,[7] let us assume that the diagnostic practices, based on a given cultural background, in fact, ground facts about anxiety disorders. In this spirit, one may come up with an explanation of the following type:

> [1] She suffers from anxiety disorder because her behavior violates a specific social norm, or shared expectation concerning the ability to speak publicly in front of strangers.

Taken in isolation, this explanation has *two* interpretations, corresponding to the two versions of social constructivism, a causal and a non-causal one. On the former, norm-violation plays a causal role in the occurrence of the anxiety disorder; and it may, in this respect, be similar to the case of evaporation of water and the human causes of global warming. On the latter, a behavioral pattern *counts* as an instance of anxiety disorder (in part), *because* it is an instance of a norm-violation.

Consider the *causal* reading first. Assume that, for some reason, a subject develops a minor anxiety concerning a particular type of social situation, say, to deliver a speech in front of strangers, in a social context where this form of anxiety, and the behavioral patterns that go together with it, are conceived of as socially awkward, or at least as not fulfilling a shared expectation. Showing the relevant behavior (say, some form of avoidance behavior, or specific behavior while delivering a speech) constitutes a norm violation; people react to the norm violation, thereby enforcing the anxiety in the subject – to a degree that it becomes pathological. In this case, actual social feedback in response to norm violation may trigger a feeling of shame, which, in turn, may cause the person to experience distress, which, in turn, may cause further deviations from socially expected behavior up to a degree that makes the condition pathological. Here, violation of social norms plays a causal role for the etiology of the disorder. This can easily be seen once we note the temporal component involved in the underlying process: *first*, there was norm violation which caused a certain behavior in the audience. The behavior in the audience *then* caused further distress, which, *after some time* and repeated stressful experiences, resulted in the development of an anxiety disorder.

On the other interpretation, the explanation does not commit one to there being a *development from* norm-violation *to* mental disorder. This is the interpretation (Stier's) Wakefield appears to have in mind, when claiming that there is a normative impact of society on the concept of a mental disorder (on my interpretation: that the social norms sometimes determine, in a conceptual or metaphysical sense, what is a disorder and what is not). On this interpretation, *norm-violation* (or *being disposed to violate certain norms*) and *suffering from anxiety disorder* occur synchronically. The former partly grounds the latter, or, put differently, the latter can be metaphysically explained in terms of the former. There are several ways in which one can cash out talk of metaphysical explanation. For instance, one may suggest that the truth of a proposition that a person has a mental disorder is explained by a truth about the violation of social norms. Or, alternatively, one may want to claim that the instantiation of the property of having a mental disorder metaphysically depends on the occurrence of norm-violations. We need not go into the details here. For our present purposes, suffice it to note that there are at least two interpretations of the explanation that a person suffers from a mental disorder due to some social facts, a causal and a non-causal one, and that whatever the correct explication of the non-causal interpretation is, it will render the explanation true without any implications concerning a possible causal (and, thus, temporal) connection between norm-violation and having a mental disorder. An understanding of the distinction between causal and metaphysical dependence along these lines is sufficiently precise for the goals of the present paper. Social constructivism may involve both, a causal and a non-causal claim concerning the relation between mental disorders and social facts. Although this is probably true of most scientific explanation, it is worth pointing out that these explanations are, of course, only *partial* explanations; the presence of the explanans phenomenon is not fully explained in terms of its social causes or some underlying social facts. Versions of constructivism we will be dealing with in what follows are claims about the *partial* social construction of mental disorders. Purely physiological, behavioral, or experiential aspects that are not themselves socially constructed may be required to offer a full explanation of the relevant phenomena of mental disorders.

Before we turn to the specific versions of social constructivism about psychiatry – shouldn't we say bit more about what makes social constructivism about psychiatry *social*? Intuitively, what depends on shared attitudes (like money), what varies with cultural context (like the social status of, say, a widow), or what itself essentially depends on a social object (like the property of playing in the NBL), is, in a sense, itself a social object. As a social object, it requires, at some stage, a sort of construction. There are straightforward examples of socially constructed objects; but this does not mean that there is a straightforward characterization of what social constructivism consists in. Any (at least partly

[7] Stier is, as it seems, not quite right when suggesting that Wakefield *thereby* supports the claim that "the cultural setup […] tends to dictate the boundary between the normal and the deviant on the basis of the expected values and virtues of its members" [(2), p. 28]. Rather, Wakefield's observation supports the epistemic claim that the perceived boundaries between the normal and the deviant are dictated (or maybe better: partly influenced) by the cultural setup. Wakefield stresses that there is a difference between false diagnostic practices and "social phobia [which is] a real disorder in which people can sometimes not engage in the most routine social interaction" [(14), p. 154].

successful) attempt to deliver a general answer to the question of what social constructivism is would transcend the boundaries of the present paper; but a general answer is not required – the theses we will be concerned with here are committed to the social dimension of their targets (such as particular mental disorders) in a *straightforward* sense; and we will be able to relate, in passing, these cases to intuitive formulations of social constructivism. Short reflection on two versions, one might, at first sight, want to classify as versions of social constructivism will help to get a better understanding of what makes social constructivism about psychiatry a version of *social* constructivism, and it will reveal that not all constructivist rhetoric amounts to constructivism as opposed to objectivism.

Versions of Social Constructivism about Psychiatry

Social constructivism about psychiatry, as introduced above in contrast with radical objectivism, is heavily underdetermined. It is underdetermined with respect to its *target*, it is underdetermined with respect to the target's *grounds*, and it is underdetermined with respect to the *relation* allegedly holding between the target and the target's grounds. In this section, I will introduce five candidate targets for social constructivism about mental disorder: mental disorders themselves, the system of symptoms, phenomenology, and experience associated with a mental disorder, the second order property of being a mental disorder, and articulation of experiences and interpretation of utterances in patient–clinician interaction. The latter two can be dismissed immediately – they invite constructivist rhetoric at best. The candidate targets will have an impact on the candidate grounds and the relevant relation supposedly holding between the two.

Consider a person who has been diagnosed with PTSD and is, based on this diagnosis, classified as suffering from a mental disorder. This will, on the side of the patient, involve (a) PTSD itself, with its specific history, including the etiology or the trigger of PTSD, (b) a specific subjective experience and, possibly, a specific phenomenology, and a set of symptoms that are indicative of PTSD. Note that depending on the view one adopts regarding the nature of mental disorders, these may ultimately collapse into one single target, if the phenomenology, the subjective experience and particular symptoms enter the individuating criteria for the disorder itself;[8] but even if this were the case, talk about symptoms, phenomenology, experience, and the disorder would still be acceptable. Drawing the terminological distinction appears to be innocent.

Furthermore, the classification of PTSD as a mental disorder will involve (c) a specific aspect of PTSD in virtue of which it counts as a mental disorder (rather than, say, a stressful episode of minor importance) and becomes clinically relevant. Finally, being diagnosed with and treated for PTSD typically requires that the patient interact with a clinician. She will (d) express her experiences and inner perspective as well as report symptoms in a particular way. On the side of the clinician, (e) an interpretation of the observed and reported (verbal and non-verbal) behavior of the patient is required.

For each of these targets, one can subscribe to the view that it is shaped by social facts. We have thus identified five candidate targets for social constructivism, all of which may give rise to a form of social constructivism about psychiatry, or at least may go together with some constructivist rhetoric. And indeed, all of these can be found in the literature. Before we turn to the more promising candidates for serious versions of social constructivism, let me briefly comment on the last two alleged targets, and corresponding claims concerning the impact of culture and social environment. Little reflection will reveal that cultural influence on patient–clinician interaction is irrelevant in the context of social constructivism about psychiatry, properly construed.

The DSM 5 contains a section "Cultural Formulation," a revised version of what had already been presented in the previous manual. Its goal is somewhat difficult to identify. It contains information on mental disorder in relation to, well, *anything culture*, so to speak, beautifully illustrated by the suggested "Overall cultural assessment":

> Summarize the implications of the components of the cultural formulation identified in earlier sections of the Outline for diagnosis and other clinically relevant issues or problems as well as appropriate management and treatment intervention. (DSM 5, 750)

Information gathered about cultural background – including religious background and, possibly, gender identity – should inform diagnosis, clinician–patient interaction, and intervention. One particular reason for an assessment of cultural background is that differences in cultural background may cause confusion and misunderstandings. So, the manual includes questions that aim, in particular, at a clarification of *cultural concepts* and *idioms of distress* (DSM 5, 758 ff.), *some* of which concern the way the individual or members of the group the individual belongs to verbalize a given experience. One contention is, in this context, that knowledge of sociocultural background may facilitate access to underlying conditions; the idea does not seem to be that cultural expression of a disorder forms an integral part of the disorder itself (we will turn back to this below, in Section "Social Context, Experience, Phenomenology, and Symptoms").

Does the claim that clinicians should be sensitive the culture-specific articulations of the underlying disorder constitute a version of social constructivism, in any interesting sense? This does not seem to be the case. To use an idiom from the natural science: some of the *data* we gather may be difficult to interpret. In psychiatry, the data may be difficult to interpret because they, first, involve verbal reports, whose real or intended meaning may escape the interpreter; and, second, the interpreter may exhibit something like a cultural bias, which makes interpretation of data

[8] Let me give just two examples of views according to which symptoms may enter the individuating criteria for psychiatric types. If you assume, with Wakefield (15), that disorders are harmful dysfunctions, and that whether or not a certain psychological condition is harmful or a dysfunction may depend on the symptoms it produces, then symptoms may enter the individuating criteria for psychiatric taxonomies. You will end up with a similar result if you subscribe to what Murphy calls the "neo-Krapelinian picture" according to which "mental illnesses are regularly co-occurring clusters of signs and symptoms that doubtless depend on physical processes but are not defined or classified in terms of those physical processes" (16).

difficult, too. But once properly interpreted, the content of the knowledge we acquire need not be knowledge involving cultural facts, just because access to such knowledge required reflection on cultural background. Let me illustrate the point by way of a (fictional) example, building on what one may want to call an informed stereotype: assume you enter a shop in Berlin, run by locals, say, a bakery. You buy a roll and a cake, adding up to € 2.45. You offer a € 20 bill. It may very well happen that the clerk looks at you like you've insulted him, refuses to take the bill, rolls his eyes and says: "Damn, I don't have any change left!" Interestingly, this really does not mean that he does not have any change left. It appears to be some form of culturally determined expression of what elsewhere would probably have been expressed by something like: "Excuse me, do you have small change?" To properly interpret the utterance, knowledge of the cultural background is required. But of course, this does not mean that the fact (that there is only little change left) is constituted by these social facts knowledge of which is required to interpret the data. Claiming that in order to properly interpret the utterance in this context you have to take cultural considerations into account, does not imply that the content of the knowledge you end up with (if you're successful) is knowledge about social facts.

Analogously, the claim that cultural considerations should play a role in clinician-patient interaction has no impact on the subject matter of psychiatry. It does have an impact on the practice of psychiatry. Data may be culturally determined. But data are not what psychiatry is ultimately about. We are mainly interested in the commitments of current psychiatry (and parts of philosophy of psychiatry) with respect to the relation between psychiatry and the natural sciences. The key question is not whether cultural differences may pose difficulties in the *assessment* of whether or not an individual suffers from a particular mental disorder. The question is, rather, whether the subject matter of psychiatry involves social explanations in terms of social facts or causes.

A similar point could be made about the recent trend to take cultural considerations into account in the context of health management [also present in the "Cultural Formulation" in DSM 5, see, for instance, questions regarding "expectations for services" that may depend on the individual's cultural background (p. 752)]. The Supplement *Mental Health: Culture, Race, and Ethnicity*, edited by he U.S. Department of Health and Human Services, calls attention to the fact that cultural or racial background may have a significant impact on access conditions to mental health institutions. Thus, reflection on cultural aspects is of utmost importance in the context of psychiatric practice. But this does not amount to social constructivism about psychiatry.

The following three sections will introduce explications of the thesis that social context and culture shape mental disorder, which amount to forms of social constructivism about psychiatry in a more demanding sense.

Social Constructivism about Mental Disorders

As noted above, one may subscribe to social constructivism concerning mental disorders themselves, such as schizophrenia, PTSD, or dissociative identity disorder. Social constructivism about mental disorders comes in two radically different forms. According to one view, which appears to be a minority view (at best), facts about mental disorders are like facts about Searlean institutions – whether or not someone is, say, an alcoholic depends, in a non-causal sense, on attitudes of other people according to which that person is an alcoholic. Being regarded as an alcoholic makes one an alcoholic. According to the other view, at least some mental disorders are individuated by their social causes, so that their instantiation non-causally depends on their etiology. This makes these disorders social in nature, although it remains to be seen whether it amounts to social constructivism, properly construed. Let me briefly comment on the first view, and then turn, in more detail, to the second.

Constructivist claims in metaphysics often take the following form:

> [2] F's are F's because they are considered/regarded to be/experienced as F's.

A famous instance of this schema is Searle's claim that "[m]oney is money because the actual participants in the institution regard it as money" [(9), p. 17]. Some formulations in statements about the nature of mental disorders allow for a similar reading. Pickering, for instance, suggests that:

> The relevant features of alcoholism do not, contrary to what it demands, exist independently of the category into which alcoholism is placed [viz. illness, *RvR*]. [(17), p. 27].

One can interpret this claim as follows: some features of alcoholism depend on specific attitudes toward alcoholism within a social context – namely, that it is regarded as alcoholism [rather than a mere moral weakness (Pickering) or, alternatively, as "manly" behavior].[9] In their critical discussion of social constructivism about mental disorders, Kendler et al. [(5), p. 1145] appear to interpret social constructivism in a similar way, associating it with Haack's characterization of a kind that is not real, i.e., not "independent of how we believe it to be" [(18), 132]. I am not entirely sure whether anyone ever held some such belief; but it appears that self-declared objectivists often tend to credit constructivists with some such view (so that the opponent of Kendler et al. would turn out to be a straw-man). One may suspect that Foucault accepted this form of constructivism (19), although Foucault, in later years, explicitly subscribed to a form of causal constructivism that is based on causal effects of discursive practices.[10] Be that as it may – on this version of social constructivism, the *targets* are categories of mental disorder, the *grounds* of the targets are attitudes or discursive practices that

[9] Pickering distances himself from a social constructivist interpretation of this view – he regards the sort of construction involved in the classification of a condition as a mental illness as non-social. On his conception of social construction, social constructivists are wedded to the idea that what is socially constructed "exists only in certain cultural or social frames of reference" [(17), p. 98].

[10] The relevant power relations involve, according to Foucault, not only attributive practices but also causal feedback, for instance, in what he called biopolitics (20).

involve the psychiatric category itself and attribute it to certain people, and the relation is that of non-causal dependence.[11]

This version of constructivism is not the only form of social constructivism regarding mental disorders. Some mental disorders seem to be *typically* caused by particular social events. PTSD is *typically* caused by experiences such as incarceration as a prisoner of war, or traumatic experiences that often have a social dimension, such as interpersonal violence or sexual assault.

Being typically brought about by social causes alone does, of course, not amount to social constructivism. To see that, consider the question of whether it contradicts radical objectivism about psychiatry. The radical objectivist may hold that mental disorders can be caused by social events, as long as we need not cite the social dimension among the relevant causes in an appropriate explanation that enables us to fully understand the mental disorder we are dealing with. The evaporation of water may often be caused by social events. A large part of the evaporation of water is currently caused by global warming (or so let us assume for the sake of the example). Global warming, in turn, is caused by our joint actions. And if global warming has a significant impact on the natural evaporation of water on earth, then, currently, the evaporation of water on earth is *typically* caused by social facts. However, this does not amount to social constructivism about the current evaporation of water – the question of whether or not evaporation of water is brought about by social causes is irrelevant to our understanding of what evaporation of water consists in. Put differently: *evaporation of water due to global warming caused by social events* is not an interesting type in physics or chemistry (though it may be an interesting type in the context of politics). The radical objectivist about mental disorders will maintain that similarly, questions regarding the causes of a disorder are, or should be, irrelevant in psychiatry.

However, trauma- and stressor-related disorders are partly *characterized* in terms of their etiology. Consider the general characterization of these types of disorder, taken from the DSM 5:

> Trauma and stressor-related disorders include disorders in which exposure to a traumatic or stressful event is listed explicitly as a diagnostic criterion. (DSM 5, 265)

If the diagnostic criteria enter the taxonomy of mental disorders, then there are mental disorders that are partly individuated by their etiology. The motivation for this relational characterization is straightforward. Different experiences cause different subtypes of trauma- or stressor-related disorders, and knowledge of the cause may bear on clinical decisions. Often, the traumatic or stressful event involves a *social* dimension. Typical examples include experience of war, torture, incarceration as a prisoner of war, or sexual abuse. DSM 5 states explicitly that "[t]he disorder may be especially severe or long-lasting when the stressor is interpersonal and intentional" (DSM 5, 275). Interpersonal and intentional stressors involve a social dimension – not because any intentional action is social, bur rather because any intentional action directed at another person – here: the patient – is social. Witnessing a death by accident of a close relative and experiencing a catastrophe such as an anaphylactic shock form the *only* exceptions in the list included in DSM 5, waking during surgery forms a borderline case (in this case, it is not clear whether the fact that surgeons appear to interfere intentionally with the patient's body plays a relevant explanatory role).

So, the family of trauma- and stressor-related disorders involves types of disorders that are, in part, defined in terms of the types of causes that actually triggered the disorder. PTSD due to traumatic experience during incarceration as a prisoner of war is different from PTSD due to traumatic experience caused by witnessing an accident involving the loss of a close relative. The difference shows at the token level, and it also shows at the level of subtypes, subtypes of the type *trauma or stressor-related disorder*, when the type of cause is referred to in a classification of the relevant disorder.

But doesn't such relational individuation of types seem odd? Upon reflection, relational classifications are common. On most accounts, *being an artwork* depends on the etiology of the artwork – namely, being intentionally produced. Similarly, being an artifact depends on the etiology of the artifact. And individual horses are, on some views, horses in part because they are descendants from horses. In the social sciences, relational types are common: someone is a president (and not merely, say, a warlord), a state (and not merely a territory), or a law (and not merely an enforced rule) only if it has a certain history.

Thus, on the view implicit in DSM 5, for some subtypes of trauma- or stressor-related disorders, if a person suffers from this subtype, she does so because of the specific etiology of the disorder. The fact that the disorder has a cause of a particular type grounds the fact that it is a disorder of the corresponding subtype (for instance, *disorder caused by an interpersonal and intentional stressor*, or *PTSD due to torture*). Expressed in terms of targets, grounds, and the alleged relation between the two: the *targets* are subtypes of trauma- and stressor-related disorders. The grounds are complex events that relevantly involve a social dimension, such as intentions directed at other persons, interpersonal actions, and social properties, such as being a war and being an incarceration, which, moreover, stand in a causal relation to the psychological condition classified as a disorder. The relation between the social grounds and the occurrences of the disorder is, thus, causal. But the relation between the disorder and the cause is not *merely* causal, it is also conceptual, or metaphysical: to instantiate one of the subtypes requires that the disorder be caused by a particular type of object. The cause not only causes the disorder, facts about the cause also ground facts about the instantiation of the disorder. This marks the difference

[11] It appears that on this view, psychiatry, conceived of as a science that deals with mental disorders, would rest on a confusion. The objects of psychiatry would turn out to be mere chimeras; psychiatry would turn out to be like a version of jurisprudence whose self-declared goal was to improve the law by empirical interventions on the minds of judges, based on the false assumption that the role judges play in the juridical system is determined by features intrinsic to judges themselves, while recognizing that typical properties relevant in the legal system are social properties. Moreover, the view appears to be at odds with the view that *suffering* correlates with the presence of conditions typically classified as disorders, or better, as (Reviewer 2) suggested, with the assumption that different types of suffering correlate with different types of disorders; so, it seems that there is more to these conditions than classification or being regarded as having a mental disorder (although this may indeed constitute part of the problem).

between the current evaporation of water, on the one hand, and the relevant types of mental disorders, on the other. Whereas, in fact, some events are evaporations of water and can be explained in terms of social events, causal explanations referring to social events are not required for an explanation of why evaporation of water is evaporation of water. Suffering from PTSD due to experience of interpersonal violence, in contrast, requires a specific causal history involving a social cause.

On this interpretation, some mental disorders are clearly *social objects* in the sense that they essentially involve a social aspect. But does this amount to social *constructivism*, in a straightforward sense? It fits Haslanger's characterization, since in defining the subtype "we must make reference to social factors" [(13), p. 98]. The classification scheme involves types that are individuated by, and whose instantiation depends on, the presence of specific social features of the causes of the disorder. In some sense, however, this does not amount to social *constructivism*: it is not the case that forms of PTSD exist in or require for their existence, as Pickering put it, "a cultural or social frame of reference" [(17), p. 98]. Not every social object, such as abusive behavior, or incarceration, is a social construction in this sense. This may speak against Haslanger's characterization or at least require elaboration on the notion of a social factor. Ultimately, this may be a verbal issue about how we want to use the term "constructivism"; what is important in the present context is that this type of classification of subtypes of disorders appears to contradict *objectivism*; in the present context, we may, then, group it with more demanding versions of social constructivism, discussed in the next sections.[12]

Social Context, Experience, Phenomenology, and Symptoms

Let us move from aspects of social construction in classification to aspects of social construction involved in experience, phenomenology, and symptoms of mental disorders. Whether constructivism about experience, phenomenology, or symptoms is compatible with robust objectivism depends on whether we take these to be constitutive of the disorder.

As we have just seen, the claim that causes of mental disorders exhibit social features alone does not amount to any interesting form of social constructivism; a claim about *metaphysical* dependence on social causes was required. This should not come as a surprise. Recall the example discussed in Section "Basic Tenets of Social Constructivism":

> [1] She suffers from anxiety disorder because her behavior violates a specific social norm, or shared expectation concerning the ability to speak publicly in front of strangers.

On its causal interpretation, according to which causal norms played a causal role in the development of the disorder, committing to [1] one does not commit to any form of social constructivism. The objectivist may maintain that disorders have social causes; and the objectivist may consistently hold that *cultural variation among disorders depends on social causes.*

As already mentioned, DSM 5 offers a guide for cultural evaluations in the "Cultural Formulation," which deals, to a significant extent, with problems in the interpretation of verbal reports; but it also deals with interpretation of experiences, calling attention to the fact that "the cultural constructs … influence how the individual experiences […] his or her symptoms or problems […]" (DSM 5, 750). As long as the subject's reports are distinct from the underlying disorder, the claim that an adequate interpretation of reports requires knowledge about sociocultural background does not amount to social constructivism. The same holds for the claim that sociocultural background shapes the experience, phenomenology, and symptoms of the disorder, as long as these are not constitutive of the disorder itself, but, rather, *signs* of the disorder. On this picture, experiences, symptoms, and phenomenology may need translation just like reports. But signs of a disorder need not play an essential role in scientific classification. As long as the disorder manifests at the physiological level, and as long as types of disorders can, at least in principle, be individuated in purely physiological terms, the objectivist can happily admit that social causes have an impact on mental disorders [see, for instance, Ref. (5)]. The radical objectivist may even accept that knowledge of cultural context is epistemically relevant for clinical practice.

From the objectivist perspective, the connection between social causes and the way a disorder manifests may involve causal connections at the level of neurodevelopment. Kirmayer and Crafa (22) have, in the spirit of social or cultural neuroscience, argued that the physiological structure underlying specific conditions is shaped by social causes:

> Culture can be seen as providing essential contexts for the development and functioning of the brain on multiple timescales: through its evolutionary history, which has involved brain–culture coevolution; across individual lifespans as biographical events are inscribed in circuitry by mechanisms of epigenetics and learning; and through ongoing influences on neural functioning by specific contexts of adaptation and performance. [(22), p. 7]

[12] Although this is not the topic of the present paper, two possible objectivist responses immediately come to mind: Maybe it is not the fact that the cause has certain social features, but rather that it is *perceived as* having these features, which is relevant to the type of disorder that we are dealing with. So, if reference to social features of causes just roughly catches what really does the explanatory work – that the cause is *perceived* as being of a particular social type – social constructivism about these types of disorders appears to be mistaken. Still, it is not entirely clear how to cash out, and account for, the relevance of perceived social causes within the radical objectivist framework. Second, one could construe this form of social constructivism as a form of pragmatic constructivism. Maybe Psychiatry, in its current form, is not as good as it gets. The types current psychiatry deals with are pragmatically adequate. If, for instance, incarceration as a prisoner of war is reliably connected with showing a specific behavioral pattern and being responsive to specific treatment, then, given our current epistemic background, individuation of subtypes of PTSD with respect to the causes makes perfect sense – for purely pragmatic reasons [according to Beebee and Sabbarton-Leary (21) fulfillment of some such condition is sufficient to count as a scientific type]. On this view, the mechanism underlying the connection between particular causes and particular disorders is still missing. Once uncovered, reference to the social cause becomes idle. The view that some mental disorders should, at the present stage of development of psychiatry, be classified in terms of their social causes is, thus, compatible with versions of in-principle reductionism about mental disorders. As a consequence, one may have reservations to group pragmatic constructivism with more robust forms of constructivism.

Again, this appears to be perfectly compatible with an objectivist stance – there is no social construction involved, if Kirmayer, Crafa, and others are right. Of course, these claims raise interesting methodological issues; for instance, if sociocultural background has a significant impact on the development of the physiology underlying mental disorders, can we dispense with descriptions in terms of sociocultural background, or does it provide heuristics that, for the time being, are indispensable in clinical contexts? To repeat: epistemic questions of this sort may arise even if we adopt objectivism about mental disorders. Unless we individuate disorders with respect to sociocultural causes, as suggested in DSM 5 for PTSD, we do not end up with a form of social constructivism.

If we do, however, we would end up with a very similar form of constructivism [as (Reviewer 2) has pointed out]: one may adopt the view that experiences, symptoms, or phenomenology are constitutive of a disorder, and that experiences, some symptoms and the phenomenology that goes together with a disorder are not individuated by their underlying physiology, but, rather, require individuation in terms of their social dimension [see Ref. (23–25) for a critical perspective]. Thornton (26) has recently pointed out that Jaspers (27), in his discussion of psychiatry (for instance, in his book "Allgemeine Psychopathologie," first published in 1913), suggested that for a great number of psychological conditions that count as a mental disorder, some form of subjective understanding is required, besides observing behavior and understanding the other as rational (if possible), for an understanding of the kind of disorder the patient suffers from. Thornton tentatively agrees that an understanding of phenomenal aspects and subjective experiences may be required in understanding the nature of the disorder.

In a somewhat similar spirit, certain experiences may be regarded as forming essential parts of a disorder. Consider the following example, taken from Stier's discussion of cultural variation in psychiatry:[13]

> A […] striking cultural difference can be found in the case of social anxiety. While in the western cultural sphere this is connected with the fear of being harmed or offended, in Japan and Korea people are in fear of harming or offending others […]. [(2), p. 29]

Now, none of these claims makes, all by itself, for *social* constructivism; highlighting the way a disorder is presented to a subject from the first person perspective need not go together with social constructivism (although it will of course pose difficulties for forms of naturalism that typically go together with the form of objectivism we are concerned with here). However, the way a condition is presented to a subject from the first person perspective may *involve* a social dimension. Consider the case of variation across cultures in anxiety disorders: the assumption is that variation should be further explained in terms of cultural background, the role an individual is supposed to play within a community, the notion of personhood, etc. If we assume that the different forms of experience in western and Japanese and Korean culture correspond to at least two different subtypes of some disorder, subtypes any psychiatric taxonomy should be sensitive to, and if we assume that there is no unified physiological type corresponding to these experiences, we end up with a robust form of social constructivism. In Haslanger's words: the subtypes are "causally constructed" in that "social factors play a causal role in bringing [them] into existence or, to some substantial extent, in [their] being the way [they are]" [(13), p. 98]. Our objectivist would oppose a view according to which genuine social experiences that are not individuated by their physiological basis are essential to classification in psychiatry. The targets are disorders, and the grounds are experiences, phenomenology, or symptoms that are individuated only with respect to a social context. The relation is, again, non-causal; the perspective of the individual is not causally connected to the disorder; it is constitutive of it.[14]

So, does *no* thesis regarding the social cause of mental disorders, all by itself, amount to social constructivism? Is all social construction non-causal, as far as psychiatry is concerned?

Consider one particular version of the claim that some disorders have (maybe among others) social causes. It is somewhat atypical. Standard versions of social constructivism regarding causal influences of the cultural background on phenomenology, symptoms, or experiences have it that cultural background does not contain any particular "information" on the disorder itself, which would explain cultural variation. For instance, the difference between social anxiety in Europe and some parts of East Asia is explained in terms of general differences regarding the concept of a person or the expectations regarding the relation between an individual and the community it lives in. In contrast, Ian Hacking, and, from a psychiatric perspective, Piper and Merskey (28) have argued that some mental disorders evolve in response to specific theories of disorders, and the expectations that go together with specific theories. Hacking writes:

> [T]here was usually only one well-defined alter; today, sixteen alters is the norm. In France, a century or so ago, cases of doubling had the symptoms then associated with florid hysteria – partial paralyses, partial anesthesia, intestinal bleeding, restricted field of vision. English cases of double consciousness were more restrained but regularly went into a trance […]
>
> Times change, and so do people. People in trouble are not more constant than anyone else. But there is more to the change in the lifestyle of multiples than the passage of time. We tend to behave in ways that are expected of us, especially by authority figures – doctors, for example. Some physicians had multiples among their patients in the 1840s, but their picture of the disorder was very

[13] Stier uses this example to back up an argument against biologism; it is not entirely clear to me how this example may affect even purely epistemic forms of reductionism, unless we buy, as a premise, that for biologism to be true, there must be *one physiological condition underlying both disorders*. But this does not seem to be required for some form or another of biologism to be true.

[14] There is, of course, also a relevant causal connection between the phenomenology or experience and the social environment of the subject, but this does not appear to be constitutive for the disorder.

> different because the doctors' expectations were different. That is an example of a very general phenomenon: the looping effect of human kinds. [(29), p. 21]

Hacking's view is clearly not that there is one type of disorder that can be expressed or even experienced in different ways due to the different models of the disorder offered in a given social environment (such as *first* multiple personality disorder, *then* dissociative disorder). Rather, the idea is that the discourse in psychiatry itself has a causal impact on the disorder the patient develops. The discourse causes the individual to adopt the dissociative identity disorder personality. The discourse, including expectations and specific norms, functions as an external cause of the disorder itself.

Some remarks on the details of Hacking's account are required in order to avoid misunderstandings. Basically, Hacking distinguishes between two different types of kinds, *indifferent* and *interactive kinds*. The former are kin to natural kinds; their objects are indifferent to our classification. Interactive kinds, however, involve causal feedback mechanisms, where the subjects that fall into the extension of an interactive kind respond to the classification, which, in turn, bears on the classification itself. This sort of causal looping effect is visible, according to Hacking, in psychiatric classification. But psychiatric classification is not purely interactive. According to Hacking, mental disorders have a physiological basis, which can be classified in terms of an indifferent kind. For what follows, I will be solely concerned with Hacking's characterization of the interactive aspect of psychiatric classification.

The target is, again, a particular type of disorder. The grounds are social events, involving interaction between patient and clinician or members of the social environment, with certain individual or shared expectations. The dependence relation is causal – the person is caused to adopt a certain pattern of experiences and behavioral traits, which, in turn, may have an impact on the classification. In contrast with the versions of causal constructivism about mental disorders previously discussed, Hacking assumes that the very attribution of a disorder contributes to the disorder (or the interactive aspect of the disorder). Interestingly, this renders true an explanation that looks, at its surface level, very much like Searle's constructivist claim about money:

> [3] Some people suffer from dissociative identity disorder because others regard them as suffering from dissociative identity disorder.

Unlike Searle's claim about money, however, [3] can be interpreted causally (and it is merely a partial explanation). It appears that there are causal versions of social constructivism in psychiatry. It is a form of constructivism not because of cultural variation, or because the disorder relates to attitudes like money relates to acceptance as money; it is a form of constructivism because it credits a conceptual practice, or discourse (that of multiple dissociative disorder) with the ability to literally create its own objects; and it does so in a way that masks the actual mechanisms that underlie the occurrence of multiple dissociative disorder. Of course, it is similar to cases where cultural background shapes the symptoms; discourse about multiple dissociative disorder is part of the cultural background. And if it has an impact on the development of dissociative disorders, including behavioral patterns, experiences, and symptoms, then in this case, cultural background shapes the disorder. But in this particular case, there is more: due to the tight connection between cultural background and disorder (i.e., the content of the discourse, which determines what it is to suffer from multiple dissociative disorder, and the disorder itself) the causal looping effect, if it occurs, ensures that that the current psychological condition instantiates the (first order) properties that define the interactive aspect of the disorder. Ignoring feedback mechanisms: The discourse causally contributes to the fact that the subject adopts the behavioral pattern others expect the subject to show; and showing the pattern is one mark of the disorder. If the interactive aspect of psychiatric classification were constitutive for facts about a subject falling under the relevant category,[15] then the instantiation of the property of having dissociative identity disorder would metaphysically depend on facts about adaptation to expectations (among other facts). In this case, causal construction and metaphysical dependence go together.

Now, finally, let me turn to what may be regarded as the most common, and, at the same time, the most challenging version of social constructivism regarding psychiatry: the thesis that the property of *being a mental disorder* is itself socially constructed.

Mental Disorder and Social Norms

In the tentative characterization of the notion of a mental disorder published in the introductory parts of DSM 5, the authors state that "[a]n expectable or culturally approved response to a common stressor […] is not a mental disorder" (DSM 5, 20). It seems that thereby, the authors intend to indicate that something is a mental disorder *only if* it involves responses to stressors that are *not* expectable or culturally approved. This reading is supported by the following passage:

> The boundaries between normality and pathology vary across cultures for specific types of behaviors. Thresholds of tolerance for specific symptoms or behaviors differ across cultures, social settings, and families. Hence, the level at which an experience becomes problematic or pathological will differ. The judgment that a given behavior is abnormal and requires clinical attention depends on cultural norms that are internalized by the individual and applied by others around them, including family members and clinicians (DSM V, 14)

In contrast with the issues raised under the heading "Cultural Formulation," this is clearly more than a purely epistemological

[15] Tsou (30) has argued that if Hacking is right, then a stable classification of mental disorders in terms of the indifferent "part" of the classification should be possible. Here, we are not so much concerned with what Hacking actually claims; rather, we can, based on the work of Hacking, identify one possible version of social constructivism about mental kinds, a version of social constructivism that captures the interactive "part" of psychiatric classification.

point about access conditions. The authors suggest that whether or not a problematic experience is pathological depends in part on the social or cultural context. This idea has played a prominent role in the anti-psychiatrist movement (31), and the question of whether normative aspects involved in diagnostic procedures are social or can be cashed out in, say, descriptive or biological terminology has attracted considerable attention (3, 4, 26, 32). Being a mental disorder is a second order property, in the sense that it is supposed to be instantiated by first order properties of psychological conditions. An alcoholic exemplifies a specific psychological property, which, in turn, is supposed to exemplify the property of being a mental disorder.[16] The significance of the property of *being a mental disorder* appears to stem from the fact that it distinguishes the clinically relevant from the clinically irrelevant (just like related concepts of health and illness). Social constructivism about this second order property states that whether or not it is exemplified by a first order psychological condition depends on the social norms within a society. The boundaries between the pathological and the non-pathological determine whether or not a condition is a mental disorder, and they are themselves partly determined by social norms. Very roughly, social norms are usually regarded as a particular type of expectations about the behavior of others (33, 34); individuals are expected to follow a norm and are expected to act in a certain way when they detect norm-violation. Only if some such structure is widespread among a society, a social norm is in place. Facts about social norms are clearly social facts. Consequently, this form of social constructivism maintains that facts about the second order property *being a mental disorder* (the target of this form of social constructivism) depend on social facts, such as the presence of social norms, and that the relation between the two is non-causal dependence – the social norms do not cause a condition to be a mental disorder; they determine, in a conceptual, or logical sense whether or not some condition counts as a mental disorder.[17] Here, in a straightforward sense, instantiation of the second order property of being a mental disorder requires a certain conceptual framework; and it is shared attitudes that ground the instantiation of the property.

CONCLUSION

Let me sum up the results. Versions of social constructivism come in three different forms. Their targets are (some or all) types of mental disorders, or the second order property of being a mental disorder. Social facts and events appear among the grounds of mental disorders, and social norms or expectations shape the boundaries of the second order property of being a mental disorder. It may even be the case that sometimes, psychiatric classification itself has an impact on the occurrence of its objects, as Hacking suggests. Note that apart from versions of social constructivism about the second order property of being a mental disorder, social constructivism about psychiatry may, and probably will be limited to *some* forms of mental disorders. What is true of trauma or stressor-related disorders need not be true of schizophrenia, or, in particular, neurodevelopmental disorders.

So, what's the consequence of social constructivism for psychiatry as a science? It is difficult to give a straightforward answer. Quite a bit will hinge on general considerations about reduction, about the nature of scientific kinds, and on the nature of explanation. In order to bypass these further questions, let me begin with an observation that immediately follows from the different versions of social constructivism described above: if social constructivism about psychiatry is true, in one of the versions described above, then the conceptual apparatus psychiatry employs involves concepts of social objects, events, or facts. Moreover, it credits the so-represented events, objects, or facts with an explanatory role, in causal as well as metaphysical respects. Causal explanations in terms of social causes would turn out to be genuine and indispensable explanations within psychiatry – indispensable in the sense that the psychiatric types are partly social in nature. Although the radical objectivist may accept that some causal explanations of the occurrences of mental disorders that cite social events as causes are true, or may even play a useful heuristic role, she will deny that such explanations figure among the set of *psychiatric* explanations, properly construed. Recall: the evaporation of water from the Ocean may be caused by global warming, which, in turn, may be caused by social events. But an explanation of the evaporation of water from the Ocean based on its social causes will not count as an explanation within physics. Similarly, on the objectivist view, the explanation of mental disorders in terms of social causes will not count as a psychiatric explanation. Social constructivism, in some of its versions, is bound to deny that. Parts of psychiatry would then move toward the social sciences, as far as their explanatory practices are concerned.

Now, there is no consensus as to how we should conceive of scientific explanation, especially in the social sciences. One may take a liberal stance, suggesting that type construction and explanation mainly serve heuristic purposes, that causal generalizations merely systematize observations, and that consequently, the metaphor of the sociocultural environment *shaping* mental disorders can be given a non-committal, pragmatic, or purely epistemic interpretation. If so, the commitments of social constructivism are relatively weak; and dispute between constructivism and objectivism would be a dispute not about the nature, but rather about the pragmatically adequate or epistemically beneficial conceptualization of mental disorders. Some such considerations may have played a role in the history of DSM – the question of where to invest time and money may, by and large, be decided on pragmatic grounds.

But there is a corresponding ontic distinction that seems to better fit the nature of the dispute between constructivists and objectivists. The constructivist raises the worry that objectivism will fail for principled reasons; it just does not get the metaphysics

[16] In a derivative sense, individuals instantiate the property of having a mental disorder, namely, when they instantiate a first order property (such as alcoholism), which exemplifies the second order property of being a mental disorder.

[17] Although Wakefield departs from a biological notion of dysfunction, he arrives at a similar form of social constructivism. Wakefield characterizes the concept of a disorder in terms of a "condition [that] causes some harm or deprivation of benefit to the person as judged by the standards of the person's culture" – what he calls the "value criterion" [(15), p. 385], which complements the biologically construed dysfunction condition for the presence of a mental disorder. Wakefield assumes that disorders require biological dysfunction, but that what counts as a normal or acceptable life depends on social or cultural context.

right and, hence, looking for mechanisms or genetic profiles alone will never deliver the desired results. Rather, we should take social background, etiology, and cultural formation into account. And the objectivist maintains that at best, considerations regarding the sociocultural environment may play a heuristic role, or constitute an obstacle in the appropriate translation from behavioral observation and the subjects' reports to the language of the underlying neural mechanisms.

Social constructivism, on its metaphysical interpretation, may have ramifications that go beyond issues that pertain to the metaphysics of science in the narrow sense. Some versions of social constructivism, concerning, for instance, gender or race categories, have been proposed not as a challenging exercise in theoretical metaphysics, but rather with a critical intention. Very roughly, the critique, whatever the details, departs from the observation that what *seems* to be a distinction, or is typically regarded as a distinction grounded in natural properties, in reality is a distinction imposed on the world *by us*, to a relevant degree. If true, this may have significant political and ethical ramifications – whether a status is natural or social bears on normative considerations. Revealing the hidden social nature of an allegedly natural category may then constitute a first step in a critical enterprise [for an explication of the idea of how social constructivisms relates to the critique of current practices, cf. Ref. (35)]. Uncovering the social nature of gender categories is important not only for general metaphysical purposes but also, and primarily, for political reasons. The idea is that if what appears to be natural turns out to be social, evaluative, and political practice should change.

We cannot go into the details here; but one may expect that similar critical or ameliorative projects will be relevant in the philosophy of psychiatry whenever evaluations based on the classification of people as having a mental disorder hinge on an objectivist interpretation of the property of having a mental disorder, where, in fact, having a mental disorder depends on social facts. If social norms determine whether or not a psychological condition belongs to the category of having a mental disorder, and if being classified as having a mental disorder forms the basis of unjust treatment by others because these others mistakenly believe the property of having a mental disorder to be natural, or objective, then uncovering the social nature of the property of being a mental disorder may bear on social practice. This is precisely what Szasz intended.

Let me close with a related observation regarding the distinction between objectivists and constructivists in psychiatry. As (Reviewer 2) has stressed, variation in the occurrence of a type depending on sociocultural background (such as the occurrence of specific gender identities) is often regarded as an indication of the fact that the occurrence of this type depends on social facts; social constructivism has been the theory of choice to account for such facts. Recall Hacking's theory about the construction of dissociative identity disorders. Cultural variation is, here, clearly an indicator for the dependence of dissociative identity disorder on social facts. It is an indicator of the latter; but does cultural variation alone support some form or another of social constructivism?

Above, I have suggested that cultural variation in mental disorders is compatible with the form of objectivism discussed here. Doesn't this show that the way I use the term "objectivist," and, hence, "constructivist," is at odds with at least one important use of these terms in the general debate on social constructivism? Here, I can merely gesture at an answer. I think that there is a difference between the pair "constructivist"/"objectivist" in the context of debates about first order mental disorders and, say, in the field of gender theory. This is not an accident; it is mainly due to the fact that the type of objectivism opposed, for instance, by feminist constructivists just lacks a counterpart in the sphere of psychiatry, because gender categories differ from categories of (first order) mental disorders in one important respect. Simplifying a lot, conceptualization in terms of gender will typically go together with an implicit naturalist conception of gender properties (sometimes described as *essentialization*). Consequently, the gender-objectivist believes gender to be part of the *nature* of an individual; being a woman is supposed to be inborn, deviations from the dichotomy are regarded as, well, non-natural and, possibly, requiring intervention. These issues simply do not arise in current day psychiatry. Conceptualizing someone as suffering from a disorder does *not* suggest a conceptualization of the person as having this property by nature.

In the context of psychiatry, *naturalist* or *essentialist* objectivism typically concerns the *second order* property of being a mental disorder; objectivists claim that it is a natural or essential property of (first order) psychological conditions, whereas constructivists maintain that classification of first order psychological conditions as a mental disorder comprises an element of construction.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This work was supported by the Volkswagen Foundation as part of the generously funded project "A Study in Explanatory Power" and by the German Research Foundation within the network "Social Functions."

REFERENCES

1. US Department of Health and Human Services, editor. *Mental Health: Culture, Race, and Ethnicity – A Supplement to Mental Health: A Report of the Surgeon General*. Rockville, MD: U.S. Department of Health and Human Services; Substance Abuse and Mental Health Services Administration; Center for Mental Health Services (2001).
2. Stier M. Normative preconditions for the assessment of mental disorder. *Front Psychol* (2013) **4**:611. doi:10.3389/fpsyg.2013.00611
3. Kendell R. The concept of disease and its implications for psychiatry. *Br J Psychiatry* (1975) **127**:305–15. doi:10.1192/bjp.127.4.305
4. Boorse C. On the distinction between disease and illness. *Philos Public Aff* (1975) **5**:49–68.
5. Kendler KS, Zachar P, Craver C. What kinds of things are psychiatric disorders? *Psychol Med* (2011) **41**:1143–50. doi:10.1017/S0033291710001844
6. Hoeltje M, Schnieder B, Steinberg A, editors. *Varieties of Dependence: Ontological Dependence, Grounding, Supervenience, Response-Dependence*. Philosophia Verlag (2013).

7. Correia F, Schnieder B, editors. *Metaphysical Grounding. Understanding the Structure of Reality*. Cambridge: Cambridge University Press (2012).
8. Searle J. *The Construction of Social Reality*. London: Penguin (1995).
9. Searle J. *Making the Social World: The Structure of Human Civilization*. Oxford: Oxford University Press (2010).
10. Tuomela R. *The Philosophy of Sociality*. Oxford: Oxford University Press (2007).
11. Epstein B. *The Ant Trap. Rebuilding the Foundations of the Social Sciences*. Oxford: Oxford University Press (2015).
12. Millett K. *Sexual Politics*. London: Granada Publishing (1971).
13. Haslanger S. Ontology and social construction. *Philos Top* (1995) **23**:95–125. doi:10.5840/philtopics19952324
14. Wakefield JC. The concept of mental disorder: diagnostic implications of the harmful dysfunction analysis. *World Psychiatry* (2007) **6**:149–56.
15. Wakefield JC. The concept of mental disorder: on the boundary between biological facts and social values. *Am Psychol* (1992) **47**:373–88.
16. Murphy D. Philosophy of psychiatry. 2012 Edition. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*. (2010). Available from: http://plato.stanford.edu/entries/psychiatry/
17. Pickering N. *The Metaphor of Mental Illness*. Oxford: Oxford University Press (2006).
18. Haack S. *Defending Science – Within Reason: Between Scientism and Cynicism*. Amherst, NY: Prometheus Books (2003).
19. Foucault M. *Histoire de la folie à l'âge classique: Folie et déraison*. Paris: Plon (1961).
20. Foucault M. *Histoire de la Sexualité (vol. 1): La Volonté de Savoir*. Paris: Gallimard (1976).
21. Beebee H, Sabbarton-Leary N. Are psychiatric kinds 'real'? *Eur J Anal Philos* (2010) **6**:11–27.
22. Kirmayer L, Crafa D. What kind of science for psychiatry? *Front Hum Neurosci* (2014) **8**:435. doi:10.3389/fnhum.2014.00435
23. Gallagher S. Delusional realities. In: Broome M, Bortolotti L, editors. *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*. Oxford: Oxford University Press (2009). p. 245–66.
24. Ghaeme S. Feeling and time: the phenomenology of mood disorders, depressive realism, and existential psychotherapy. *Schizophr Bull* (2007) **33**:122–30. doi:10.1093/schbul/sbl061
25. Murphy D. *Psychiatry in the Scientific Image*. Cambridge: MIT Press (2006).
26. Thornton T. *Essential Philosophy of Psychiatry*. Oxford: Oxford University Press (2007).
27. Jaspers K. *Allgemeine Psychopathologie. Für Studierende, Ärzte und Psychologen*. Berlin: Springer (1913).
28. Piper A, Merskey H. The persistence of folly: a critical examination of dissociative identity disorder. Part I. The excesses of an improbable concept. *Can J Psychiatry* (2004) **49**:592–600.
29. Hacking I. *Rewriting the Soul*. Princeton, NJ: Princeton University Press (1995).
30. Tsou J. Hacking on the looping effects of psychiatric classifications: what is an interactive and indifferent kind? *Int Stud Philos Sci* (2007) **21**(3):329–44. doi:10.1080/02698590701589601
31. Szasz T. *The Myth of Mental Illness*. London: Paladin (1972).
32. Fulford K. Analytic philosophy, brain science and the concept of disorder. In: Bloch S, Chodoff P, Green SA, editors. *Psychiatric Ethics*. Oxford: Oxford University Press (1999). p. 161–92.
33. Elster J. *The Cement of Society*. Cambridge: Cambridge University Press (1989).
34. Bicchieri C. *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press (2006).
35. Haslanger S. Gender and race: (what) are they? (what) do we want them to be? *Noûs* (2000) **34**:31–55.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

MINI REVIEW
published: 23 November 2015
doi: 10.3389/fpsyt.2015.00166

Beyond Scientism and Skepticism: An Integrative Approach to Global Mental Health

Dan J. Stein[1] and Judy Illes[2]*

[1] MRC Unit on Anxiety and Stress Disorders, Department of Psychiatry, Groote Schuur Hospital, University of Cape Town, Cape Town, South Africa, [2] National Core for Neuroethics, Division of Neurology, Department of Medicine, University of British Columbia, Vancouver, BC, Canada

The global burden of disorders has shifted from infectious disease to non-communicable diseases, including neuropsychiatric disorders. Whereas infectious disease can sometimes be combated by targeting single causal mechanisms, such as prevention of contact-spread illness by handwashing, in the case of mental disorders multiple causal mechanisms are typically relevant. The emergent field of global mental health has emphasized the magnitude of the treatment gap, particularly in the low- and middle-income world and has paid particular attention to upstream causal factors, for example, poverty, inequality, and gender discrimination in the pathogenesis of mental disorders. However, this field has also been criticized for relying erroneously on Western paradigms of mental illness, which may not be relevant or appropriate to the low- and middle-income context. Here, it is important to steer a path between scientism and skepticism. Scientism regards mental disorders as essential categories, and takes a covering law approach to causality; skepticism regards mental disorders as merely social constructions and emphasizes the role of political power in causal relations. We propose an integrative model that emphasizes the contribution of a broad range of causal mechanisms operating at biological and societal levels to mental disorders and the consequent importance of broad spectrum and multipronged approaches to intervention.

Keywords: neuroethics, global mental health, scientism, skepticism, causal mechanisms

OPEN ACCESS

Edited by:
Annemarie Kalis, Utrecht University, Netherlands

Reviewed by:
Oksana Sorokina, The University of Edinburgh, UK
Eleftheria Pervolaraki, University of Leeds, UK

****Correspondence:***
Dan J. Stein
dan.stein@uct.ac.za

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry

Received: *10 August 2015*
Accepted: *05 November 2015*
Published: *23 November 2015*

Citation:
Stein DJ and Illes J (2015) Beyond Scientism and Skepticism: An Integrative Approach to Global Mental Health. Front. Psychiatry 6:166. doi: 10.3389/fpsyt.2015.00166

INTRODUCTION

In recent decades, there has been a shift from infectious disease to non-communicable diseases throughout the world. Mental, neurological, and substance use disorders are already the largest contributor to the burden of disease; these prevalent, chronic, and costly disorders now account for 22% of disablilty adjusted life years (DALYs) from all medical causes in those aged 15–49 (1). Furthermore, forecasts indicate that in the foreseeable future they will become even more central to global public health, with the World Economic Forum predicting that neuropsychiatric disorders will comprise the largest costs of chronic, non-communicable diseases globally in the next two decades (2).

Increased recognition of the burden of neuropsychiatric disorders has given impetus to the emergence of the discipline of global mental health (3). Additional key considerations are that interventions for mental disorders impact positively on individual well-being and country development

and are highly cost-efficient, but that neuropsychiatric disorders are often underdiagnosed and undertreated, with the treatment gap particularly large in low- and middle-income countries. Furthermore, this treatment gap is a human rights issue; levels of stigmatization of people living with mental illness are too high, and levels of mental health literacy are too low in communities, clinicians, and policy makers (4, 5).

Clinical and research work on neuropsychiatric disorders raises a number of conceptual and ethical questions, many of which are relevant to the field of global mental health (6). In considering some of these conceptual and ethical questions, we have argued that clinicians and researchers should steer a course between a scientism that regards mental disorders as natural kinds, and a skepticism that views all mental disorders as mere sociocultural constructions (7–9). Integrative approaches are needed to address fully the complex reality of mental disorders. In this commentary, we discuss this view in relation to global mental health, considering in turn issues of diagnosis, pathogenesis, and intervention (**Table 1**).

GLOBAL MENTAL HEALTH AND DIAGNOSIS

Global mental health has emphasized that mental disorders are prevalent and associated with significant suffering, impairment, and socioeconomic costs. Thus, for example, data from the World Mental Health Surveys have emphasized that mental disorders are more impairing than physical disorders, but are less likely to be diagnosed and treated (10). While such conclusions are pertinent around the globe, in low- and middle-income countries, a lack of resources is particularly likely to exacerbate the treatment gap. These sorts of data provide an important foundation for the rallying cry of global mental health that there is no health without mental health (11).

Nevertheless, global mental health has also come under fire for its emphasis on these sorts of data. In particular, critics have argued that the field relies erroneously on Western paradigms of mental illness, which may not be relevant or appropriate to the low- and middle-income context (12). Such constructs run the risk of ignoring how symptoms vary from time to time and from place to place, and of downplaying the complex ways in which illnesses are expressed and experienced differently in different sociocultural contexts (13). Indeed, Jacob and Patel have emphasized that global mental health needs new diagnostic approaches, a view that is perhaps partially consistent with attempts in clinical neuroscience to reformulate approaches to evaluation of neuropsychiatric disorders (14, 15).

At the same time, international classification systems have significant clinical advantages, and there are currently no viable alternatives in practice. We would, therefore, argue that although it is clearly important to recognize the limitations of current psychiatry nosology and biopsychosocial models (16, 17), we ought to be wary of unrealistic expectations of such approaches (18). For example, medicine does not require that its diagnostic systems are essentialist in nature; rather medical syndromes provide clinicians with a practical set of tools for assessing patients. Rather than insisting that assessment systems will ultimately be supported solely by data on endophenotypes – intermediate phenotypes with high heritability – we can also ask that more work is also needed on exophenotypes, such as societal, structural, and other upstream contributors to disease and illness (19), and their intersections.

GLOBAL MENTAL HEALTH AND PATHOGENESIS

Although infectious diseases may involve a range of biological and psychosocial factors, it is sometimes possible to combat these conditions by targeting single causal mechanisms. Locating the geographic source of a cholera epidemic, employing handwashing to decrease bacterial transmission, developing vaccines to prevent polio and smallpox, and using mosquito nets to prevent malaria have been seminal exemplars of success for public health. In contrast, global mental health has had to contend with multiple upstream factors that impact mental disorders: poverty, inequality, gender discrimination, and more.

At the same time, any emphasis of global mental health on only one set of causal factors can potentially be problematic.

TABLE 1 | Moving beyond scientism and skepticism in global mental health: integrative approaches to diagnosis, pathogenesis, and intervention.

	Scientism	Skepticism	Integrative
Diagnosis	Diagnostic systems rely on essentialist categories or natural kinds. Assessment systems will be ultimately be supported by data on endophenotypes	Mental illness is expressed and experienced differently in different sociocultural contexts. Symptoms vary from time to time and place to place	Mental illness is a complex reality. Nosologies are theory bound and value laden, but may improve as the relevant science and debate advance
Pathogenesis	May approach causality in terms of covering laws. May focus on a single set of associations, such as those which characterize the health care system	May emphasize the role of sociocultural values and powers in explanations. May focus on differences in conceptualization of disorders across history and geography	Emphasizes that a broad range of factors are involved in the pathogenesis of mental disorders, with causal mechanisms operating at multiple interacting levels
Intervention	May take a single-bullet approach, looking for focused interventions, whether biological or community focused that will target the essence of the disorder	May emphasize that interventions reflect local values and powers. Both biological and community-focused interventions reinforce existing societal structures	Incorporates a range of insights about the nature of mental disorders, and targets a broad range of factors involved in their pathogenesis, including biological and social ones

Some research priority setting exercises have indicated that global mental health should focus primarily on health systems research, for example, and should pay less attention to the biological causes of mental disorders (20). This is consistent with a criticism of global mental health which emphasizes that it is ironic that a field that purports to be concerned with a broad range of socioeconomic factors relies on neuroessentialist DSM-5 categories. After all, key considerations for Western-based typologies of illness are that they have diagnostic validity, that disorders demonstrate high heritability, or that they predict response to interventions such as pharmacotherapy.

Our own view is that there are important opportunities at the intersection of global mental health and clinical neuroscience in addressing the pathogenesis of mental disorders (21). There has been significant progress on understanding how nature and nurture intersect to create vulnerabilities for mental disorder, and indeed in recognizing how multiple levels of causal factors contribute to these conditions (7, 22). We have previously noted, for example, that while basic neuroscience has shed a great deal of insight into addiction, a full understanding of substance use disorders requires the psychological and social levels to be included (8, 9). Only a comprehensive and integrative perspective will allow an understanding of complex phenomena, such as decreased voluntary control in addictive disorders (23).

GLOBAL MENTAL HEALTH AND INTERVENTION

Global mental health has focused on task shifting and implementation science. This is certainly important in the context of resource-limited settings, where there are simply not enough trained professionals to deliver interventions, where health systems have systemic problems, and where there is growing evidence that non-specialized community workers can make a real impact (3). Indeed, some of the concerns of global mental health mirror those of the solution-oriented bent of neuroethics; there is a focus on efforts to improve wellness, on the importance of human rights, and on an empirical approach to optimizing interventions (6).

At the same time, there are potential criticisms of the focus of global mental health on communities, task shifting, and implementation science. Sartorius and colleagues, for example, have noted that in many parts of the globe, communities have changed in significant ways and are no longer able to provide the support that those with serious mental illness need and deserve (24). Furthermore, some tasks simply cannot be shifted, and we need to focus at times rather on novel biological treatments (25) or on increasing resources; it is crucial that in attempting to strengthen resource-limited systems, we do not simply institutionalize mechanisms that can only work in impoverished systems.

Again, we would argue for an integrative approach. It is important to avoid a scientism which states that given that mental disorders are natural kinds, they will ultimately succumb to single-bullet biological interventions (26). At the same time, we do not want to fall to prey to a skepticism that indicates that interventions should be entirely focused on changing the way in which disorders are conceptualized and labeled by society, or that they should be limited to community practices. We need an integrative approach to intervention that incorporates a range of insights about the nature of mental disorders and that targets a broad range of factors involved in their pathogenesis, including psychobiological factors and community processes.

CONCLUSION

We have argued elsewhere that it is important to avoid neuroreductionism and to emphasize instead that mental and substance use disorders require an understanding of psychosocial factors. Put differently, it is important to steer a path between scientism, which regards mental disorders as essential categories and takes a covering law approach to causality, and skepticism, which regards mental disorders merely as social constructions and reduces causality to considerations of political power. Here, we have applied these arguments to the newly emergent field of global mental health, considering issues relevant to diagnosis, pathogenesis, and treatment (**Table 1**), and emphasizing that a broad range of causal mechanisms operating at biological, psychological, and societal levels, and at the interactions between these levels, contribute to mental disorders, and that clinical interventions and research practices must match this complexity.

FUNDING

DS is supported by the Medical Research Council of South Africa. JI is Canada Research Chair in Neuroethics.

REFERENCES

1. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* (2012) **380**(9859):2197–223. doi:10.1016/S0140-6736(12)61689-4
2. Bloom DEC, et al. *The Global Economic Burden of Non-Communicable Diseases*. Geneva: World Economic Forum (2011).
3. Patel V. Global mental health: from science to action. *Harv Rev Psychiatry* (2012) **20**(1):6–12. doi:10.3109/10673229.2012.649108
4. Ganasen KA, Parker S, Hugo CJ, Stein DJ, Emsley RA, Seedat S. Mental health literacy: focus on developing countries. *Afr J Psychiatry (Johannesbg)* (2008) **11**(1):23–8.
5. Semrau M, Evans-Lacko S, Koschorke M, Ashenafi L, Thornicroft G. Stigma and discrimination related to mental illness in low- and middle-income countries. *Epidemiol Psychiatr Sci* (2015) **24**(5):382–94. doi:10.1017/S2045796015000359
6. Stein DJ, Giordano J. Global mental health and neuroethics. *BMC Med* (2015) **13**(1):274. doi:10.1186/s12916-015-0509-y
7. Stein DJ. *Philosophy of Psychopharmacology*. Cambridge: Cambridge University Press (2008).
8. Buchman DZ, Skinner W, Illes J. Negotiating the relationship between addiction, ethics, and brain science. *AJOB Neurosci* (2010) **1**(1):36–45. doi:10.1080/21507740.2010.490168
9. Di Pietro N, Illes J, Canadian Working Group on Antipsychotic Medications and Children. Rising antipsychotic prescriptions for children and youth:

cross-sectoral solutions for a multimodal problem. *CMAJ* (2014) **186**(9):653–4. doi:10.1503/cmaj.131604
10. Wang PS, Angermeyer M, Borges G, Bruffaerts R, Tat Chiu W, De Girolamo G, et al. Delay and failure in treatment seeking after first onset of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry* (2007) **6**(3):177–85.
11. Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips MR, et al. No health without mental health. *Lancet* (2007) **370**(9590):859–77. doi:10.1016/S0140-6736(07)61238-0
12. Summerfield D. Afterword: against 'global mental health'. *Transcult Psychiatry* (2012) **49**(3–4):519–30. doi:10.1177/1363461512454701
13. Stein DJ. Cross-cultural psychiatry and the DSM-IV. *Compr Psychiatry* (1993) **34**(5):322–9. doi:10.1016/0010-440X(93)90018-Y
14. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med* (2013) **11**:126. doi:10.1186/1741-7015-11-126
15. Jacob KS, Patel V. Classification of mental disorders: a global mental health perspective. *Lancet* (2014) **383**(9926):1433–5. doi:10.1016/S0140-6736(13)62382-X
16. McLaren N. A critical review of the biopsychosocial model. *Aust N Z J Psychiatry* (1998) **32**(1):86–92;discussion93–86. doi:10.3109/00048679809062712
17. Ghaemi SN. The rise and fall of the biopsychosocial model. *Br J Psychiatry* (2009) **195**(1):3–4. doi:10.1192/bjp.bp.109.063859
18. Nesse RM, Stein DJ. Towards a genuinely medical model for psychiatric nosology. *BMC Med* (2012) **10**:5. doi:10.1186/1741-7015-10-5
19. Stein DJ, Lund C, Nesse RM. Classification systems in psychiatry: diagnosis and global mental health in the era of DSM-5 and ICD-11. *Curr Opin Psychiatry* (2013) **26**(5):493–7. doi:10.1097/YCO.0b013e3283642dfd
20. Tomlinson M, Rudan I, Saxena S, Swartz L, Tsai AC, Patel V. Setting priorities for global mental health research. *Bull World Health Organ* (2009) **87**(6):438–46. doi:10.2471/BLT.08.054353
21. Stein DJ, He Y, Phillips A, Sahakian BJ, Williams J, Patel V. Global mental health and neuroscience: potential synergies. *Lancet Psychiatry* (2015) **2**:178–85. doi:10.1016/S2215-0366(15)00014-0
22. Kendler KS. The dappled nature of causes of psychiatric illness: replacing the organic-functional/hardware-software dichotomy with empirically based pluralism. *Mol Psychiatry* (2012) **17**(4):377–88. doi:10.1038/mp.2011.153
23. Stein DJ. Philosophy of psychopharmacology. *Perspect Biol Med* (1998) **41**(2):200–11. doi:10.1353/pbm.1998.0037
24. Volpe U, Mihai A, Jordanova V, Sartorius N. The pathways to mental healthcare worldwide: a systematic review. *Curr Opin Psychiatry* (2015) **28**(4):299–306. doi:10.1097/YCO.0000000000000164
25. Collins PY, Patel V, Joestl SS, March D, Insel TR, Daar AS, et al. Grand challenges in global mental health. *Nature* (2011) **475**(7354):27–30. doi:10.1038/475027a
26. Stein DJ. Psychopharmacology and natural kinds: a conceptual framework. In: Kincaid H, Sullivan JA, editors. *Classifying Psychopathology: Mental Kinds and Natural Kinds*. Cambridge, MA: MIT Press (2014). p. 65–74.

Conflict of Interest Statement: In the past 3 years, DS has received research grants and/or consultancy honoraria from AMBRF, Biocodex, Cipla, Lundbeck, National Responsible Gambling Foundation, Novartis, Servier, and Sun. JI has no conflicts to declare.

frontiers in Psychiatry

HYPOTHESIS AND THEORY
published: 20 November 2015
doi: 10.3389/fpsyt.2015.00164

Causality in Psychiatry: A Hybrid Symptom Network Construct Model

*Gerald Young**

York University, Toronto, ON, Canada

Causality or etiology in psychiatry is marked by standard biomedical, reductionistic models (symptoms reflect the construct involved) that inform approaches to nosology, or classification, such as in the DSM-5 [Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; (1)]. However, network approaches to symptom interaction [i.e., symptoms are formative of the construct; e.g., (2), for posttraumatic stress disorder (PTSD)] are being developed that speak to bottom-up processes in mental disorder, in contrast to the typical top-down psychological construct approach. The present article presents a hybrid top-down, bottom-up model of the relationship between symptoms and mental disorder, viewing symptom expression and their causal complex as a reciprocally dynamic system with multiple levels, from lower-order symptoms in interaction to higher-order constructs affecting them. The hybrid model hinges on good understanding of systems theory in which it is embedded, so that the article reviews in depth non-linear dynamical systems theory (NLDST). The article applies the concept of emergent circular causality (3) to symptom development, as well. Conclusions consider that symptoms vary over several dimensions, including: subjectivity; objectivity; conscious motivation effort; and unconscious influences, and the degree to which individual (e.g., meaning) and universal (e.g., causal) processes are involved. The opposition between science and skepticism is a complex one that the article addresses in final comments.

Keywords: causality, symptom, mental disorder, construct, network

OPEN ACCESS

Edited by:
Annemarie Kalis, Utrecht University, Netherlands

Reviewed by:
Stijn Vanheule, Ghent University, Belgium
Riet Van Bork, University of Amsterdam, Netherlands

***Correspondence:**
Gerald Young
gyoung@glendon.yorku.ca

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry

Received: *23 September 2015*
Accepted: *30 October 2015*
Published: *20 November 2015*

Citation:
Young G (2015) Causality in Psychiatry: A Hybrid Symptom Network Construct Model. Front. Psychiatry 6:164. doi: 10.3389/fpsyt.2015.00164

CAUSALITY IN PSYCHIATRY: A HYBRID SYMPTOM NETWORK CONSTRUCT MODEL

The article tackles fundamental issues in psychiatry while proposing novel solutions. In particular, it considers the relationship between symptoms and disorder by examining extant models and current research. It attempts to disambiguate some of the confusions related to understanding and researching the models, preparing the way for presentation of a genuinely hybrid one based on systems theory thinking. Moreover, it presents other novel concepts related to emergent causality, and the relationship of meaning and causality in symptoms (hermeneutic insight and causal explanation; Verstehen, Erklären, respectively). The article presents a complex view of the relationship between meaning and causality involving three dimensions.

INTRODUCTION

Opposing Models

The article reflects on two types of models, a latent variable model (or construct model), which is seen as a top-down approach to understanding the relationship between symptoms and disorders, and a symptom interaction model (or network model), which is seen as a bottom-up approach to understanding the relationship between symptoms and disorders. In latent variable modeling, an underlying construct (e.g., depression) is considered causal of the relationship of the items or behaviors (e.g., symptoms) that are subsumed by the variable. In an item or behavior interaction or networked model (e.g., symptoms), relatively few direct relations are considered causal of the item/behavior/symptom relationships, which are deemed to lie among the latter themselves.

In the first model of the two involved, which is the traditional approach, symptoms (items) reflect a common underlying psychological construct and, therefore, this type of model is considered "reflective" (4). In this construct model, the cause of the mental symptom/disorder derives from the central construct, whether a disorder or a cluster, downward to the symptoms. In the symptom-interactive model, symptoms (items) mutually affect each other, and can be represented by a composite variable, but the direction of the causality is from the symptom interactions to the composite. The model is referred to as "formative" (4).

In this network model, which is the second of the two involved, causality springs from the symptoms (or clusters) interacting among themselves, a process that acts to change the symptoms/clusters (or initiate them). The composite variable is involved only as representation.

Before describing the hybrid model in depth, some of the challenges in doing so are described. This leads to presentation of a literature review preparatory to it.

Systems

The article will consider the following crucial questions. First, what do we miss when we represent disorders solely with top-down models such as the construct model? What do we miss when we represent disorders solely with bottom-up models such as a network model? In order to answer these questions, the hybrid model that has been created is framed in Non-Linear Dynamical Systems Theory (NLDST), which can be viewed both as a model that is an umbrella one or superordinate one to the construct and network ones. Therefore, the article presents a novel hybrid model, which combines these two types of models (top-down, bottom-up) into a framework that both respects them yet adds to them without detracting from them.

In this work, researchers might obtain a covariance matrix related to the multiple symptoms in a study (referring to the covariance among scores of participants with respect to the symptoms that were measured). Once the matrix is established, the covariance obtained could be explained from either a common construct perspective or from that of symptom network interactions (i.e., common cause vs. direct causal relations). In this regard, the researcher evaluates either (a) the shared variance of all measured variables of a putative construct, e.g., estimating factor loadings, and the causal pattern is from the construct to the variables; or (b) the parameters for the direct relations between symptoms. Furthermore, in one type of hybrid approach, the variance that is not explained by the common construct might be explained residually through direct relations between networked symptoms.

That being said, the present hybrid model is not built on statistical synergies but conceptual ones. It presents a theoretically plausible causal model and the statistical task, then becomes to fit extant statistical approaches to the model or expand them for this purpose. The conceptual hybrid model is built on NLDST, and the multilevel hierarchical structure that it includes allows for upper levels of the system to work with lower levels in establishing the system whole. That is, if we equate psychological constructs with emergent higher-order system levels that might derive from lower-order levels and their bottom-up interactions, such as in networks, then the stage is set for having higher-order levels reciprocally influence in turn in a top-down fashion the networks involved. For example, depression might not only be constructed by its symptoms but also it might exist as a subjective mental content or disorder and influence the configuration of its symptoms (in context, and for the individual in her/his uniqueness).

If one excludes psychological constructs from consideration as a higher-order level in a systems model, the hybrid model as presented will be dismissed. However, if one allows for its inclusion in a systems framework, as described, the framework can readily be conceived as one that has emergent higher-order levels (or constructs, e.g., mental content and disorder) that can interact top-down and reciprocally with symptom networks in their bottom-up influence on the system.

Clusters

Another complication in developing a hybrid model involving construct and network approaches to symptom–disorder relationships involves clusters, which stand intermediate between symptoms and disorders. In network modeling, subsets (clusters) of items, behaviors, or symptoms (variables) might be found, but they are not considered as independent sources of causation relative to the direct relationships among the variables. Rather, variables within any one cluster might causally influence each other in their network. Inter-variable correlations will result, but they would not reflect the causal influence of a common underpinning construct. The article will deal with this issue as it proceeds in creating a genuine hybrid model.

On the one hand, the DSM-5 [Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; (1)] includes many disorders that, through its polythetic approach to symptom identification, involve symptom clusters. However, the research on how many clusters are needed in the DSM's disorders stands as an ongoing enterprise. For example, in posttraumatic stress disorder (PTSD), the empirical findings on how to group its symptoms keeps finding an increasing amount of clusters. The number of factors or dimensions involved in research on PTSD has moved it from the DSM-5's four-dimensional model to ones with even seven and eight dimensions (see below). On the other hand, in systems modeling, there is no reason why intermediate levels cannot constitute both top-down causal levels working on lower ones and levels that can be influenced by those lower ones,

while having levels superordinate to them influence them, while they form networks among themselves that can influence their superordinate levels. Therefore, systems modeling can accommodate the concept of clusters in symptoms.

The final prefatory note about the present hybrid model of symptom–mental disorder relations (with its mutual bottom-up and top-down influences) is that its hybrid nature is not synonymous with an attempt to explain everything related to the question to the point that its inclusivity can really explain nothing. In this regard, there are many multifactorial models in causal explanation that are acceptable; there are many systems models in this regard; and there are many advancing conceptual and statistical notions that are hybrid and explanatory without being obtuse and untestable [e.g., Ref. (5, 6)].

ISSUES

In the following, the first substantive section of the article considers relevant concepts and terms. This lays the foundation for the literature review, model building, and applications.

Concepts

Psychiatry has been criticized at multiple levels, including its difficulties with its diagnostic manuals and their assumptions. It embraces mostly the medical model of disorder and diagnosis, the biocentric model of the causality of disorder, or etiology, and the psychopharmacological model of disorder treatment and management (6–9). Even the basic concept of what constitutes a mental disorder has been disputed. In the following, I review aspects of these issues, preparing the way for presentation of my own work in the area. The section ends with an integrated view of what is mental disorder.

The RDoC project [Research Domain of Criteria; (10, 11)] contends that it offers a broad approach to causation in psychiatry, but its critics maintain that it is especially biomedical, neurocentric, and reductionistic [e.g., Ref. (6, 12, 13)]. Similarly, the DSM-5 is a psychiatric classificatory system that aims to include reliable and valid categories of mental disorder with clear causes (etiology), an aspiration that, if realized, would facilitate effective treatment; however, its critics maintain that it fails to achieve its objective [e.g., Ref. (6–9)]. Also, in terms of causal explanations, they maintain that it is still steeped in the biocentric model.

Multifactorial causal models in psychiatry have been formulated, such as the biopsychosocial model [e.g., Ref. (5, 6)]. Moreover, newer modeling efforts are specifying the mechanisms in the interactions among causal influences on behavior and its disturbance, such as work on networks [e.g., Ref. (2, 14)] and attractor dynamics [e.g., Ref. (15)].

To highlight in more depth the main argument of the article, bottom-up causality in psychiatry refers to the interaction and mutual influence of symptoms in mental illness, while top-down ones refer to the influence of underlying latent variables or constructs on symptom expression. A genuinely interactive bottom-up, top-down model would acknowledge both the reality of an underlying latent variable or construct in influencing symptomatology and also networked symptom connections as influencing the underlying construct.

This approach might be antithetical to those who hold either a network or construct view of system–disorder relations, but there are advantages to the model. Moreover, it fits the overarching model of systems theory. In this regard, the next section of the article explains in depth the concept of systems, which includes different levels, self-organization, emergence, and attractors.

Non-Linear Dynamical Systems Theory

This section of the article reviews some critical concepts in NLDST. Detailed presentation of systems theory is beyond the scope of the present work; the reader should consult Thelen and Smith (16); Young (3), and also Bielczyk et al. (15).

NLDST is distinguished by its emphasis on self-organized emergence in system component interactions within and across levels. In particular, higher-order levels of systems might emerge through bottom-up interactive processes. For example, Vallacher et al. (17) referred to the emergence of "global properties" or "coherent higher-order" states through the adjustment to each other of the individual system elements involved in a bottom to top (bottom-up, instead of top-down) self-organizational process. Typically, self-organization does not reach the new system end-state instantaneously. Rather, there are many ongoing mutual system element adjustments that take place.

Through its concepts of emergence and self-organization, NLDST allows for explanation of how higher-order patterns in behavior, from the simplest limb movements to the most profound thoughts, are part of the species' repertoire. New systems states that emerge in a system function to constrain behavior emanating from the system. New state system input transforms toward state characteristics even if they are discrepant with them. That is, systems maintain stability once formed, even if perturbed, until further mutual element and input interactions lead to critical state transition points.

System states might change over time, but when they consistently return to the same state after perturbation, the state involved is considered an attractor. Attractors reside in landscapes with basins; and the wider are the basins, the more likely a range of states in the system will converge on one attractor, which metaphorically could be considered to reside at the bottom of the basin involved. In this model, the "deeper" is the basin, the greater is the attractor's resistance to perturbation.

When a system has two or more states, it is considered multistable. The attractors could involve negative or undesirable states, such as having in the same person antagonism in conjunction with antagonism avoidance. Or, the two members of a couple could be living antagonistic regimes [e.g., Ref. (18)]. Beyond attractors on which system dynamics converge, an attractive force could be like a "repellor," or one that scrupulously avoids regions in its state space rather than returning to it. Metaphorically, instead of in a basin, a repellor resides on top of a hill in the system's landscape.

Systems might have no attractors, and therefore be more susceptible to external influences. Or, systems might have one attractor, sustain a perturbation that is critical (19), and produce self (re)organization through the effect on set points in control parameters in the system.

Finally for Vallacher et al. (17), dynamic properties can be found at "different levels of psychological reality," and dynamic transformations can take place at different time scales (seconds, years). Also, network concepts fall under the rubric of dynamic ones. That is, network nodes represent elements in systems. This notion is important for present purposes in that it justifies considering network models as part of larger ones in NLDST that includes higher-order levels that can be represented as constructs.

Samuelson et al. (20) emphasized the relevance of emergence in NLDST. Emergence takes place through systems components that interact and mutually influence each other in a soft-assembly process, or from the ground, rather than from pre-specified central, top-down (deterministic) explicit coding or organization. It takes place over multiple time scales; can happen on the moment; and is conditioned by context and the history of the organism, so that the outcome is unique and variable. Systems might also have subsystems, which are strongly coupled or integrated components that are only weakly coupled at best to other components. The authors give the example of seeking hidden objects in a first location even after viewing its hiding in a second one. Research shows that, in infants, the error involved (A-not-B) is a product of cognitive and motor components in interaction, with temporal and neural dynamics at work, too.

Hayes et al. (21) noted that dynamic systems concern pattern formation and change. The principles in dynamic systems science cut across biology, ecology, political science, and other disciplines, including physics and chemistry. Systems are adaptive when they maintain a dynamic tension between stability and variability. Although resilience to perturbation can be beneficial, it should not be overly rigid. For example, from a network perspective, in depression, negative emotions exhibit stronger temporal connections (22). Psychotherapy can help in shifting maladaptive connections to adaptive ones, as demonstrated in the research of Hayes and colleagues and Schiepek et al. (23).

NLDST is a mathematical model that is conducive to psychological theorizing. For example, attractors can be represented by mathematical formalisms, and state spaces or trajectories in a system can be represented by graphical representation of differential equations (24). In this regard, modeling could include approaches such as dynamic factor analysis and application of ergodic theory (25). Butner et al. (24) explained that, mathematically, Lyapunov exponents represent the strength of system topological features, for example, the rate a system changes toward or away from a particular state (the basin steepness). They can be calculated locally (e.g., at a set point) or globally (for the whole system).

Rabinovich et al. (26) described dynamic transformation as allowing cognition and mind to emerge from brain and computation. Cognition is not reflected in any one brain center or even in the entire brain, but is a product of interconnected cooperativity over many elements. The spatiotemporal patterns in brain dynamics that are highly coherent could be called modes, and they reflect the play of extinction and stabilizing inhibition. Brain center clusters that form in tasks represent dynamical modes and correspond to transient system states. Superordinate levels in systems can be conceived as hierarchically arranged chunking networks, e.g., from sentences to paragraphs to chapters in texts.

Wichers et al. (27) indicated how moment-to-moment affect dynamics can be viewed from NLDST. They referred to research showing that symptom networks in (severe) psychopathology are more strongly interconnected than those of people with less severe psychopathology [e.g., Ref. (14)]. The networks exhibit vicious circles because their nodes reinforce each other. In dynamical terms, a system could be "very" stable and therefore even "strong" perturbations might not create variability, let alone a small one allowing for "critical transition" to another state at a "tipping point." However, in psychopathology, if there is high, mutually reinforcing connectivity with networks, such that vicious circles develop in the background without being noticed, the mood system could become fragile and vulnerable to transition, even when one node (affective state) is triggered for the reason that others are also activated in the network. A cascade effect results that continues to resonate in the network such that the "little" perturbation of the one node involved leads to a disproportionate mood change or critical transition (as in the well-known butterfly effect).

Mental Disorder and Symptoms

Before continuing in the article with the literature review and detailed modeling, the concept of a symptom needs clarification. This is undertaken toward better understanding mental disorder, which is also discussed in this section.

Symptom

A symptom is defied as a physical or mental feature that is a departure from a typical state or feeling and that might be indicative of a disease, disturbance, disorder, unusual state, or condition (and which might be noticed by the patient). Symptoms might be subjectively experienced and phenomenologically reported or objectively obtained (signs, e.g., in laboratory tests). Symptoms include the contents of mental states that might recursively influence other symptoms, such as through the vicious circles that take place after catastrophic thinking and fears. Beliefs are powerful engines driving symptomatology, as are moods, affect, emotions, drives, desires, and motivation.

Much of the work in psychotherapy relates to these mental contents, the narratives people tell about themselves and to others, and the meanings ascribed to events, as well as one's past experiences, and one's place in the present and future, in addition to other people's actions and reactions to the person and their relationships with the person. In this sense, the person is as much, if not more, a seat of the causality of symptomatology experienced as are biological (nature) and environmental (nurture) factors (as per the biopsychosocial/biopersonalsocial models mentioned previously). That being said, not all symptoms can be taken at face value or are even genuine. This is especially true because of the influence of unconscious processes on symptom expression, as well as even conscious ones, such as those related to feigning or malingering for monetary gain, as might happen with PTSD, the exemplar chosen in the article. This difference between patient presentation and actual symptomatology constitutes the quandary confronting clinicians, as well as the challenge that they and their patients must work through.

Mental Disorder

As for defining mental disorder, there is no one accepted definition. The approach of the DSM-5 involves a clinically significant disturbance reflecting dysfunction usually associated with distress or disability in activity. In contrast, for the DSM, normally neither an expected, culturally approved loss to a common stressor/loss nor individual-society conflicts, involving socially deviant behavior, are representative of mental disorder.

The DSM-5's definition of mental disorder includes an error in reasoning (8). It indicates tautologically that a mental disorder is caused by a disturbance in mental functioning, which simply uses the same words on both sides of the definitional equation. The World Health Organization (28) definition of mental disorder does not help resolve the matter. It refers to a disorder as a combination of thoughts, perceptions, emotions, behavior, and relationships that are "abnormal."

Closer inspection of the DSM-5 definition of mental disorder indicates that it is constituted by different levels. The DSM-5 definition has implicit in it several levels. They include mental function atop the hierarchy, then mental disorder as one branch. The collection of signs (objective) and symptoms (subjective) happen behaviorally, emotionally (in regulation) and cognitively, and together constitute a syndrome. Furthermore, the ensemble of signs and symptoms are associated with a "clinically" significant disturbance and usually a "significant distress, or disability." Finally, at another level implicating causality, there are "dysfunctions" in psychological, biological, or developmental "processes." Aside from these levels, often, mental disorder in the DSM-5 includes clusters of symptoms intermediate between the disorder and symptom list.

Both the DSM-5 definition of mental disorder and the WHO's definition do not include directly environment, support or its lack, or context. A relational and systemic approach to mental disorder might better arrive at its acceptable and inclusive definition [e.g., Ref. (8)].

Young (6) developed a more elaborative definition of mental disorder. According to him, it involves "a behavioral syndrome (or pattern or network of symptoms) in context that is characterized as a clinically significant disturbance, distress, or dysfunction potentially evaluated as harmful to the individual, to others, or to both." To establish clinical significance, well-informed (and trained) individuals should rely on reliable and relevant evidence. The mental disorder can be expressed in cognition, mood, relations, interactions, self-regulation, and other behavior and its organization. Biological, social, and personal (i.e., psychological), as well as developmental processes, might be factors. Social, occupational, or other important functional activities might be involved in impairment, and they might meet disability thresholds. The definition of mental disorder that I have provided is based on the DSM's approach, but broadens it, for example, by mentioning symptom networks, which is important in the present context.

Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

According to Vanheule (9), a diagnostic category needs to be both reliable and valid. For reliability, he noted that the range of kappa results in reliability studies of DSM categories has shifted in the qualitative attribute given to the best results, the next best, and so on. In particular, Clarke et al. (29) used a shift in describing kappa results that seemingly allowed for acceptable reliability for quite a few DSM-5 categories in the DSM-5 field trials, when use of the kappa ranges used in research on prior versions of the DSM would have shown questionable reliability for those DSM-5 results had the prior adjectives in summarizing kappa results had been applied without change. That being said, I note that the best results were obtained for PTSD (along with a few others; PTSD reliability results for the DSM-5 were considered at least fair or very good, depending on the criteria).

As for validity, Vanheule (9) queried whether the DSM-5 accounts well for context and whether its categories apply well to individual cases. He concluded that, rather than symptoms being signs or indices, they should be conceived as personal constructions.

The next section of the article examines recent literature related to topics in mental disorder. They include work on network models and the construct approach at issue in the article, preparing the way for the hybrid model developed over the two approaches. In brief, the articles cited have helped lead to the present top-down (construct)/bottom-up (symptom network) causal model relating symptom and mental disorder. In addition, the review provides comments that prepare elaboration of the present model.

LITERATURE REVIEW

The literature review concentrates on the disorders of posttraumatic stress and depression, in particular. It especially analyzes the research by McNally et al. (2) for the former, and Wigman et al. (14) for the latter.

Posttraumatic Stress Disorder

Dimensions

In a literature review and conceptual analysis, Rosen and Lilienfeld (30) evaluated the core assumptions of PTSD. They found that research findings provided no compelling or consistent support for its core assumptions. They queried whether it is a diagnostic category that should be kept in the DSM. In a later publication, Rosen and colleagues called for a process of active questioning to determine its validity (31). That being said, the literature has consistently engaged in scientific investigation of PTSD and its validity. Previously, I noted that the DSM-5 field trials found it to be reliable. In the following, I examine one aspect of its validity – concerning its symptom structure. The review will show that, rather than the current four-cluster model for PTSD in the DSM-5, models with more factors better fit the 20-symptom symptom list for PTSD as found in the DSM-5. In particular, a seven-factor model has been found to be the most powerful and, moreover, it has been found to have associations indicative of its clinical and theoretical value (32, 33).

Table 1 indicates the basic symptoms of PTSD both in the psychiatric diagnostic (nosological) manual, the DSM-IV [Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; (34)] and the DSM-5, and how the symptoms of PTSD

TABLE 1 | DSM-5 PTSD symptom cluster model (seven) and the dissociative subtype (one), with hypothesized core/non-core symptoms specified.

Number	Symptom	Cluster	Non-core	Core
PTSD				
1	Memories (intrusive)	Re-experiencing	–	✓
2	Nightmares (recurrent)	Re-experiencing	✓	–
3	Dissociative reactions/ flashbacks	Re-experiencing	✓	–
4	Emotional reactivity (heightened; to signals)	Re-experiencing	✓	–
5	Physiological reactivity to reminders (marked)	Re-experiencing	✓	–
6	Avoid thoughts/feelings/ memories (reminders)	Avoidance	✓	–
7	Avoid external reminders	Avoidance	–	✓
8	Amnesia: inability to recall important aspects	Negative affect	✓	–
9	Negative beliefs (persistent, heightened)	Negative affect	✓	–
10	Self/other blame (persistent)	Negative affect	✓	–
11	Negative emotional state (persistent)	Negative affect	–	✓
12	Loss of interest (marked)	Anhedonia	✓	–
13	Detachment	Anhedonia	–	✓
14	Restricted positive affect	Anhedonia	✓	–
15	Irritability/anger	Externalizing behavior	–	✓
16	Reckless/ self-destructive	Externalizing behavior	✓	–
17	Hypervigilance	Alterations in arousal and reactivity	✓	–
18	Startle (exaggerated)	Alterations in arousal and reactivity	–	✓
19	Difficulty concentrating	Dysphoric arousal	✓	–
20	Sleep disturbance	Dysphoric arousal	–	✓
Dissociative subtype				
1	Depersonalization	Dissociation	–	✓
2	Derealization	Dissociation	✓	–

The table indicates the 20 symptoms in the DSM-5 [Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; (1)], and the 17 in the DSM-IV [Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; (34)]. There have been changes in wording for some of the symptoms from one version to the next, and the abbreviated versions in the table refer to the DSM-5 symptom list. One of the symptoms in the table (sense of foreshortened future) applies only to the DSM-IV. The table also gives the arrangement of the symptoms into clusters in the DSM-5 and in the DSM-IV. There are four clusters for the former and three for the latter. The DSM-5 clusters essentially involve splitting the DSM-IV avoidance/numbing cluster, consistent with the factor analytic model on the DSM-IV factor structure (37). There are other competing models. Moreover, other research (38) points to a separate cluster related to the DSM's dissociative subtype. The table indicates which of the symptoms for each of the eight clusters in the table appear to be predominant, essential, or core (39). Adapted from Young (36).

are organized into clusters in these manuals. Moreover, the factor analytic research on how the symptoms cluster have not supported the way the DSMs have parsed the PTSD symptoms into clusters. In this regard, as mentioned, the most recent research of PTSD symptom clustering has indicated that a seven-factor model fits best how the 20 symptoms of PTSD in the DSM-5 organize into clusters [Ref. (32); and replicated by Wang et al. (35); as summarized in Ref. (36)]. Furthermore, the research supports a dissociative PTSD subtype. In this regard, I have argued that there are really eight dimensions to consider in PTSD as described in the DSM-5 (36). Finally, the tables show my approach to which are core symptoms rather than non-core ones among the PTSD symptoms in each of the eight clusters involved in PTSD and the dissociative subtype in the DSM-5 (36).

The factor analytic research on PTSD uses confirmatory factor analysis (CFA), which is based on *a priori* models that are tested. Until recently, only four-factor models had been supported, but the work of Elhai et al. (40) had shown that there might be five factors involved in PTSD, and more recently two six-factor models were tested and supported (32, 41) before they were combined in the seven-factor model. Therefore, the understanding of PTSD is becoming more refined and each cluster found represents some psychological construct related to it (e.g., re-experiencing, avoidance, hyperarousal/reactivity). This research is valuable for differentiating models of PTSD at the level of the higher-order constructs that comprise it, because working with 17 to 20 symptoms or so is quite difficult clinically. This is one reason why I tried to isolate the core symptoms in each of the clusters involved.

Zelazny and Simms (42) conducted CFA in a study of psychiatric outpatients assessed using DSM-5 symptom criteria of PTSD. For both samples studied (those meeting either criteria in interview or a subthreshold stressor), the best fit of the data involved the above-mentioned seven-factor model. However, a new six-factor model also fit well the data (named alternate dysphoria, in which difficult concentrating and sleep problems are removed from the dysphoric arousal factor in these models and placed in the dysphoria factor).

Clearly, the research continues on the factor structure of PTSD. That being said, the lack of final response to the question cannot be taken to invalidate PTSD.

Networks

Partly in reaction to the complexity of working with long lists of symptoms, researchers using the symptom network approach to PTSD are attempting to discern how symptoms coordinate into nodes and their relations, referred to as edges. Also, they seek the centrality of symptoms in networks, such as in measures of betweenness. The approach statistically is quite different than that of CFA, which focuses on underlying constructs. In network approaches, the nodes and edges are the foci, and symptom themselves in their networking create and influence each other outside of any putative underlying construct.

In the network approach, symptoms covary, or couple variably, and affect each other through feedback loops, homeostatic relations, and so on, allowing sensitivity to individual differences in symptom expression and their causality. For example, an episode of PTSD would follow a course related to symptom nodes in the network "turning on" and "transmitting their activation" to nodes connected to them.

McNally et al. (2) presented a network approach to the symptoms of PTSD. They conducted a questionnaire study of

survivors of a 2008 Chinese earthquake, with over 360 respondents. They used a translated version of the PCL [Posttraumatic Checklist – Civilian; (43); Mandarin Chinese version; (44)]. The questionnaire is keyed to the DSM-IV. According to the questionnaire, 38% met the criteria for probable PTSD (5 years after the earthquake when the data were gathered).

The data showed that with exclusion of results at $r \leq 0.30$, strong associations become more evident, for example, for hypervigilance and startle and also avoidance of thoughts and activities (about the trauma and associated with it, respectively). Numbing and dissociation symptoms were strongly linked (loss of interest in enjoyable activities; feeling distance from others, respectively). Finally, nightmares, flashbacks, and intrusive memories related to the trauma were tightly linked. The authors noted that these various symptom linkages appear related to the three DSM-IV symptom clusters of hyperarousal, avoidance/numbing, and re-experiencing, respectively. However, other symptom linkages did not conform to these DSM clusters – those of startle-concentration problems, and anger-concentration problems.

Other results included that concentration networking indicated that two re-experiencing symptoms were not connected to the others (physiological reactivity, feeling upset at reminders), but quite connected to each other. Centrality calculations showed that a highly central symptom concerns perceiving the future as foreshortened. Overall, the authors concluded that hypervigilance, future foreshortening, and sleep appear predominant symptoms in PTSD symptom network analysis, with multiple symptom linkages involved, including some not previously considered.

To conclude this section of the paper, I note that in Young et al. (45), I attempted to show how a network model of PTSD symptoms could distinguish primary (core), secondary, and tertiary ones. That work indicates that network thinking can be applied to mental disorder in multiple ways.

Depression and Other Disorders

Bielczyk et al. (15) adopted a similar model for major depressive disorder. According to them, causal relations in network dynamics are the cause of clinical constructs such as depression.

Bielczyk et al. (15) added a role in depression of attractor dynamics and also for the regulation of excitation–inhibition balance across brain circuits. These latter concepts are quite consistent with my own (6), in that I argue that NLDST can help explain shifts to health and illness attractors and that activation/inhibition coordination is an important mechanism at all levels in brain–behavior relations.

Conway and Kovacs (46) have shown how the field of human intelligence is moving away from the traditional latent psychological construct model (*g*, general factor of intelligence) in which *g* is considered a causal general ability, to new models that interpret *g* as an emergent property reflecting the positive correlations found among test scores. This research shows how the concept that underlying constructs need to be complemented if not replaced by other models is gaining traction in areas other than psychopathology.

These newer models are "formative" ones, and not the traditional "reflective," essentialist, or "entity realism" ones. In formative models, there still are psychological constructs, but as causal effects or consequences rather than causal initiators.

Conway and Kovacs (46) concluded that hybrid models of intelligence exist, and they are partly reflective in nature and partly formative, too, such as found in their own "process overlap" theory. As has been emphasized, for the topic of psychopathology, the present work also is proposing a hybrid reflective (top-down) and formative (bottom-up) causal model of the relationship between symptom and illness. The model that I have created derives from the seminal work of McNally et al. (2) and also that of Wigman et al. (14), presented next.

Wigman et al. (14) examined data gathered by experience sampling methodology (ESM) in a pooled sample ($N = 599$) of three groups (depression in past; current status mild; psychotic symptoms, with disorder diagnosed; controls). Participants were given wristwatches that beeped quasi-randomly 10 times per day over a period of 5 to 6 days (depending on the particular sample). The signal required them to fill in a self-assessment diary. The focus of their study was to analyze the relations of participants' responses, as given on a 7-point Likert scale, for five items, which were – at this moment, I feel: cheerful; content; insecure; down; suspicious.

Wigman et al. (14) reviewed the top-down psychological construct approach to mental disorder. As shown, in this approach, mental symptoms are viewed as being caused by underlying constructs. In contrast, the bottom-up approach that they reviewed maintains that psychopathology involves a complex interacting network of components. At the symptom level, this approach views mental states as nodes that, when activated, might trigger other mental states (47). Symptom networks might be non-linear in their mutual effects, reciprocal, with feedback loops, vicious circles, and increased connectivity.

Despite pointing out the major differences in the two models of how symptoms and disorder might relate, Wigman et al. (14) did not contrast in a direct fashion one model vs. the other. Rather, they compared mental state network structure over groups having the different diagnoses mentioned to healthy controls. Also, they sought clusters of network components across network data. Note that the network characteristics analyzed involved centrality indices: node strength, outward degree, inward degree, closeness, and betweenness.

Perhaps because the first type of analysis undertaken involved group comparison and the second transdiagnostic analysis, the authors referred to the cross-group network analysis as top-down even though it makes more sense to refer to network analysis as bottom-up and they referred to the principle component analysis as bottom-up even though it typically would be referred to as top-down compared to network analysis. In short, I query whether their approach by Wigman et al. (14) allows for the hybrid reflective–formative conceptualization of mental disorder and their relations to symptoms. It seems that all they did was analyze the data involved with the two types of statistics typically associated with one approach or the other, but not in the way that the statistics are typically used in this type of research. Careful analysis of their results in what follows confirms this impression.

The results in Wigman et al. (14) showed that having a diagnosis led to more strongly connected moment-to-moment

mental state network structures, and more so for depression relative to psychosis. For example, in depressed patients, there were many more interconnections between negative and positive emotions, unlike the case for the group with psychosis, for which connections like this were rare. In the latter group, there appears to be two separate loops of mental state, one negative and the other positive. In terms of the connectedness measurement, it was higher in the group with depression, e.g., in terms of node strength and inward and outward degree. Depressed individuals had the highest node strength. Finally, the comparison group had the least connections going to or coming from negative mental states.

As for the principal component analysis results, seven high-order components emerged. They were based on loadings over associations such as mental state at time $t - 1$ and what follows at time t. For the first factor, all the loaded associations began with a positive emotional state at time $t - 1$. Therefore, the authors interpreted it "impact of positive mental state." The second factor seemed to reflect the negative impact of feeling down on other mental states, and so on. A primary result was that compared to the controls, the two psychiatrically disordered groups obtained higher scores on the component of "impact of insecure," suggesting that this component might be a general one in mental illness in multiple dimensions of psychopathology. The authors noted that the network paradigm appears to be a useful one in mapping transdiagnostic processes in mental state.

Wigman et al. (14) concluded that individuals with the same diagnosis might exhibit substantially different symptom patterns. Moreover, the concept itself of separate diagnoses might be problematic in that psychopathology might reflect one underlying explanatory principle – that of mental state interconnectivity underlying symptoms. These conclusions are quite accurate, but they might reflect the manner in which the analyses were conducted rather than anything like a genuinely hybrid model of reflective and formative models. In such an approach, psychological construct and network analyses would be conceptualized as equal and interacting reciprocal causality mechanisms of the relationship of symptom and psychological construct, and not be considered hybrid simply because a principal component analysis was applied to network data. In the following, I attempt to create such an integrative model of symptom–mental disorder relations. In the latter approach, only the data analysis methods are hybrid, not the conceptualization.

This penultimate section of the article follows next and outlines a genuine hybrid model of symptom–mental disorder relations from a bottom-up–top-down perspective. Once the model building is complete, the article considers further the nature of symptoms, for example, in terms of the value of perceiving them as individualized mental content and meaning.

THE HYBRID SYMPTOM–MENTAL DISORDER MODEL

Specifically for this section of the article, I re-introduce the bottom-up and top-down models of the symptom–mental disorder relationship. Then, I show how the two models can reciprocally interrelate.

Modeling

In models of symptoms and mental disorder relations, one set of models concerns higher-order (latent variable) constructs (e.g., PTSD) that cause or influence in a top-down manner the lower-order manifest symptoms and their clusters (which in turn might be an intermediate level of influence on symptoms). In contrast, according to network models, cluster/symptom interactions cause their pattern of expressions and the term associated with mental disorder (e.g., PTSD) is a representation of the symptoms and their interactions rather than being a causal influence on their manifestation.

One way of accommodating the different views on psychopathology of what constitutes bottom-up and top-down processes is to consider a systems model with different levels (see **Figure 1**). Systems thinking is best exemplified by NLDST [e.g., Ref. (3, 16)], in which different levels of a system might interact and even be created in the interaction, just as different elements in any one level of the system (or of the system as a whole) might interact and even create new elements. Moreover, changes in system state might take place because of minor perturbations when the system is at far-from-equilibrium, while living systems generally might be poised at this state preparatory to change because of its adaptive value.

Figure 1 illustrates the difference between a hybrid conceptualization of symptom–mental state relations and a hybrid statistical analysis of the relations. There is no reason why it cannot be the case that with each of the classic top-down, psychological construct approach and the bottom-up, symptom-driven approach, there are both network and cluster statistics that could be used.

Although powerful, the network approach to symptom causality and connection is more descriptive than mechanistic. It might indicate that symptoms connect and even in unique ways compared to conjectures and findings based on other approaches. However, the causes of the connectivities involved are not specified except by indicating that the symptom interactions are the cause. This might represent tautology, although I am sympathetic to the lower-order, grounded, and micromoment dynamics producing the connectivities. We need to know which symptoms emerge as predominant in any one moment of time, and the work of Wigman et al. (14) provides methods for tackling this issue. Moreover, they also refer to dynamic temporal processes in network expression.

Nevertheless, at another level, systems theory could tell us more about emerging connectivities over symptoms and their relations to mental state. In this regard, the work of Bielczyk et al. (15) indicates that mechanisms that might cohere symptoms (or repel them) might act through dynamical system processes [including activation/inhibition balancing, which is a concept central in my work: (3, 6)]. Symptom system dynamics can be measured in different ways in dynamical systems approaches compared to network ones, for example, in terms of control and order parameters and of exponents related to bifurcation points in which systems split into new attractor regimes or chaotic–antichaotic adaptive systems, fractal patterns, and so on.

That being said, the micromoment approach to symptom connectivity at times $t - 1$, t, $t + 1$, etc., could inform these analyses in complementary ways. For example, patients might

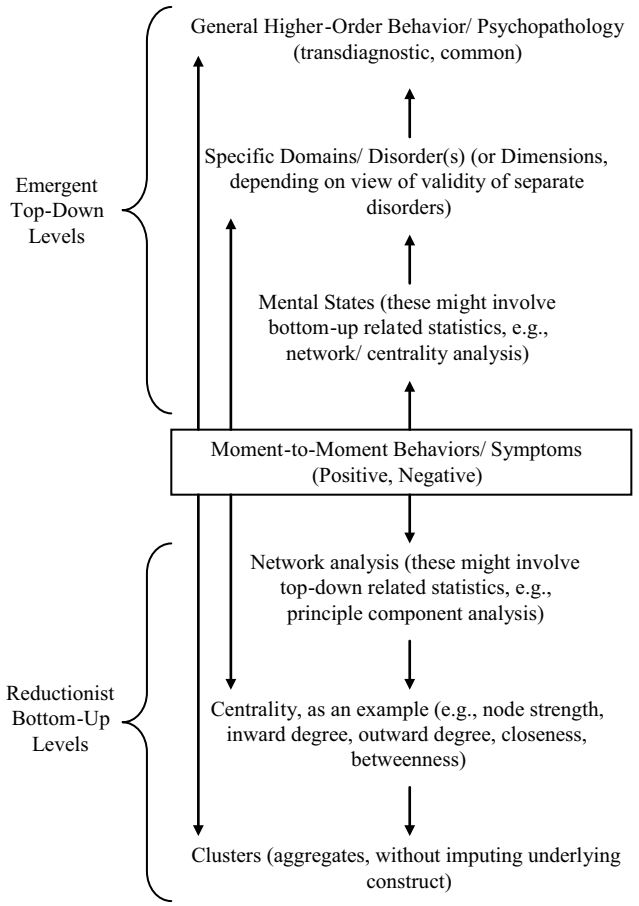


FIGURE 1 | Top-down and bottom-up levels in behavior/symptom normality/psychopathology. The figure illustrates the complementarity of top-down and bottom-up models of the causal relationship between symptoms and mental disorder. In systems, different levels might exist with higher-order levels emergent or distinct in characteristics from lower-order ones on which they are built. In terms of the constituents of the levels, they might be single elements, patterns, or networks, or even superordinate structures. Statistical techniques, such as network analysis or principle component analysis, can be applied to both emergent, higher-order levels and more basic ones. In terms of psychological constructs, at the higher-order emergent levels, in psychopathology, they might be specific disorders, dimensions, or overarching constructs (e.g., internalizing/ externalizing disorders).

have a more powerful symptom at any one time among their suite of symptoms, or one symptom might lead the way at any one moment in bringing a subthreshold one to disorder (and perhaps disability). As yet, there is no clear integrative model of how any one symptom might become primary in these senses at any one moment, although, as shown, the work of Wichers et al. (27) has made strides in these regards.

The symptom complex of the patient is crucial, as are symptom linkages over individualized patterns, or the network of nodes/ edges (relations) expressed by the patient over time. Based on this approach, the clinician might develop individual mappings of the dynamic evolution of symptoms over sessions and apply individualized approaches to intervention and treatment.

To conclude this portion of the article, hybrid conceptualizations to date on the relationship of symptoms and disorder have much to offer, but there might be conceptual limitations in the work that bar further progress. In this regard, current hybrid models [e.g., Ref. (14)], as argued above, might not allow for genuine reciprocity between the causal effects of the higher-order construct and the lower-order symptoms. Only by avoiding to equate any statistical modeling with conceptual ones and also by finding a common conceptual umbrella for both types of models can a genuine hybrid one over them be constructed. The next section presents a systems model-informed hybrid model of symptom–mental disorder relations based on these premises.

The Model

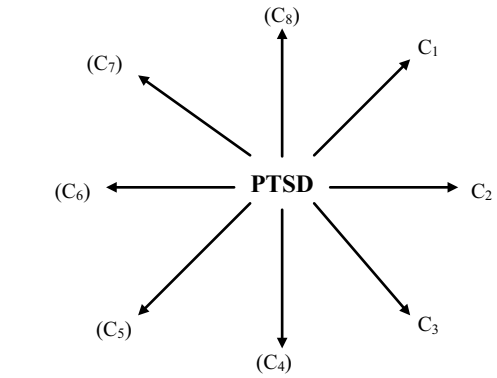
Figures 2 and **3** present core material of the present model of how symptoms and mental disorder interrelate in a hybrid fashion. The second of these two figures specifies how the concept of emergent circular causality (3) can be applied equally to the bottom-up and top-down approaches to causality. Specifically, **Figures 2** and **3** depict the difference between the latent variable/psychological construct model of the relationship between PTSD and its clusters/symptoms and the symptom-interactive or network model.

In considering development of a genuine hybrid model over the construct and symptom network approaches to how symptoms and mental disorder relates such that construct and symptoms causally interact, primacy should not be given to either component. Moreover, the statistical models that one might choose to work within each paradigm constrain the model building involved.

The next section of the article specifically demonstrates how a more integrative model of the reflective construct and formative network models could be constructed for the question of symptom–mental disorder relationship. It avoids some of the pitfalls of prior attempts to do the same. Nevertheless, it is an initial conceptualization that itself has limitations, such as not yet being mathematically grounded nor empirically tested.

Figures 4 and **5** present a genuine hybrid reflective and formative model of causality over mental symptom and disorder. For any one construct or cluster, there is not only influence/creation downward to symptoms but also feedback upward from symptom interactions to construct/cluster. Moreover, these top-down and bottom-up models function at multiple intermediary levels (intermediate, superordinate) and the interactions can take place not only horizontally (among symptoms; among levels/sublevels; and their configurations/patterns) but also vertically (downward or upward over (sub) levels).

Therefore, causality does not reside in one nexus node, level, element, element (sub)set, construct, or multiple aspects of these constituents of the symptom and disorder but in all the rich dynamical systemic interactions and reciprocal influences among them. Symptoms have causal effects on each other but constructs have causal effects on them. Constructs, such as mental disorder, are not ephemeral, reducible entities to symptoms, but emergent, irreducible entities that can affect and even initiate the symptoms. They reflect dynamical system characteristics, and can take on a life of their own at higher-order levels of a system. Perhaps they are not directly observable, but their role



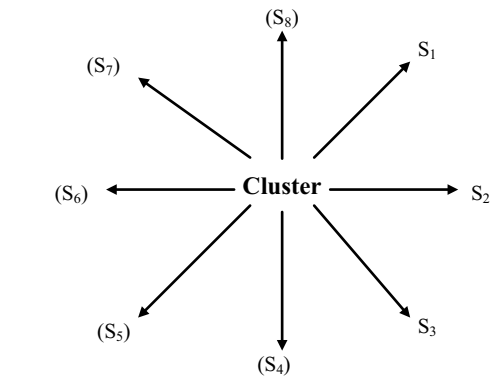
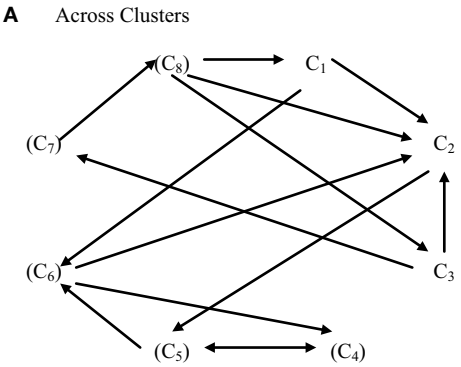


FIGURE 2 | A latent variable construct (top-down) causal model of PTSD symptoms (S) and clusters (C). **(A)** Across clusters, **(B)** within clusters. In latent variable or construct models of psychological phenomena, an "essential" underlying psychological entity, trait, characteristic, or superordinate attribute is considered as a valid higher-order behavioral reality that is not caused by or conditioned by the lower-order behaviors/symptoms associated with it but, to the contrary, conditions or causes in a top-down manner how they are manifested (in context, over time/development). Mental disorders might have several clusters and each can be characterized as a quasi-dependent sub-disorder that conditions/causes its associated symptoms. In this model, individual differences derive from the overarching construct involved and not from the manifested symptoms themselves, which merely reflect, in their patterns, the higher-order individual differences involved.



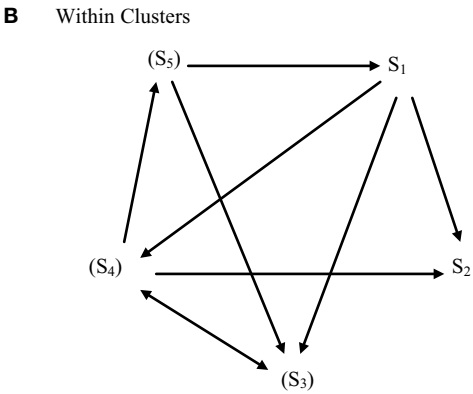


FIGURE 3 | A symptom-interactive (bottom-up) causal model of PTSD symptoms (S) and clusters (C). **(A)** Across clusters, **(B)** within clusters. In "non-essentialist" system-interactive or behavior/symptom network, connective models, behaviors/symptoms interact amongst themselves and constitute the cause of the pattern of behaviors/symptoms expressed. For example, if sleep is poor, other symptoms might be exacerbated. Individual differences in behavior/symptom expression derive from the behavior/symptom interactions in context (and over time/development). There is no higher-order "essential" (latent) psychological variable, construct, entity, trait, characteristic, or attribute that influences the behavior/symptom interactions. If terms relating to these levels of behavior are used in this model, it is only to represent the interactions and not as a factor that causes or influences them. In this regard, behaviors/symptoms in interaction do so at a level that is bottom-up rather than top-down.

can be inferred and the mechanisms that being them about are increasingly understood.

In short, emergence is a common construct in systems theory, but in my approach to it, circular causality constitutes an important driving mechanism in emergence (48). That is, as system levels interact with one another, new ones can emerge at higher orders, and they can become overarching and overriding drivers of behavior and symptom expression (3, 6). Specifically, I had written in Young (3) that in "circular emergence" different levels of systems can form and integrate, with higher-order ones gaining degrees of freedom through their flexibility even as their degrees of freedom are constrained through the intercoordinations involved. Also, I noted that activation/inhibition coordination can serve as the critical mechanism in stabilizing systems, in keeping them at the cusp of change, and in recreating equilibrium after they change.

CONCLUSION

First, the article has provided background information, such as relevant definitions and issues related to nosology, causality, and network and construct models of symptom–disorder relations. Then, it reviewed the relevant literature in the field, tackling alternative models and trying to disambiguate them. Next, it gave a genuinely hybrid model for the relationship of symptom network and psychological constructs in mental disorder.

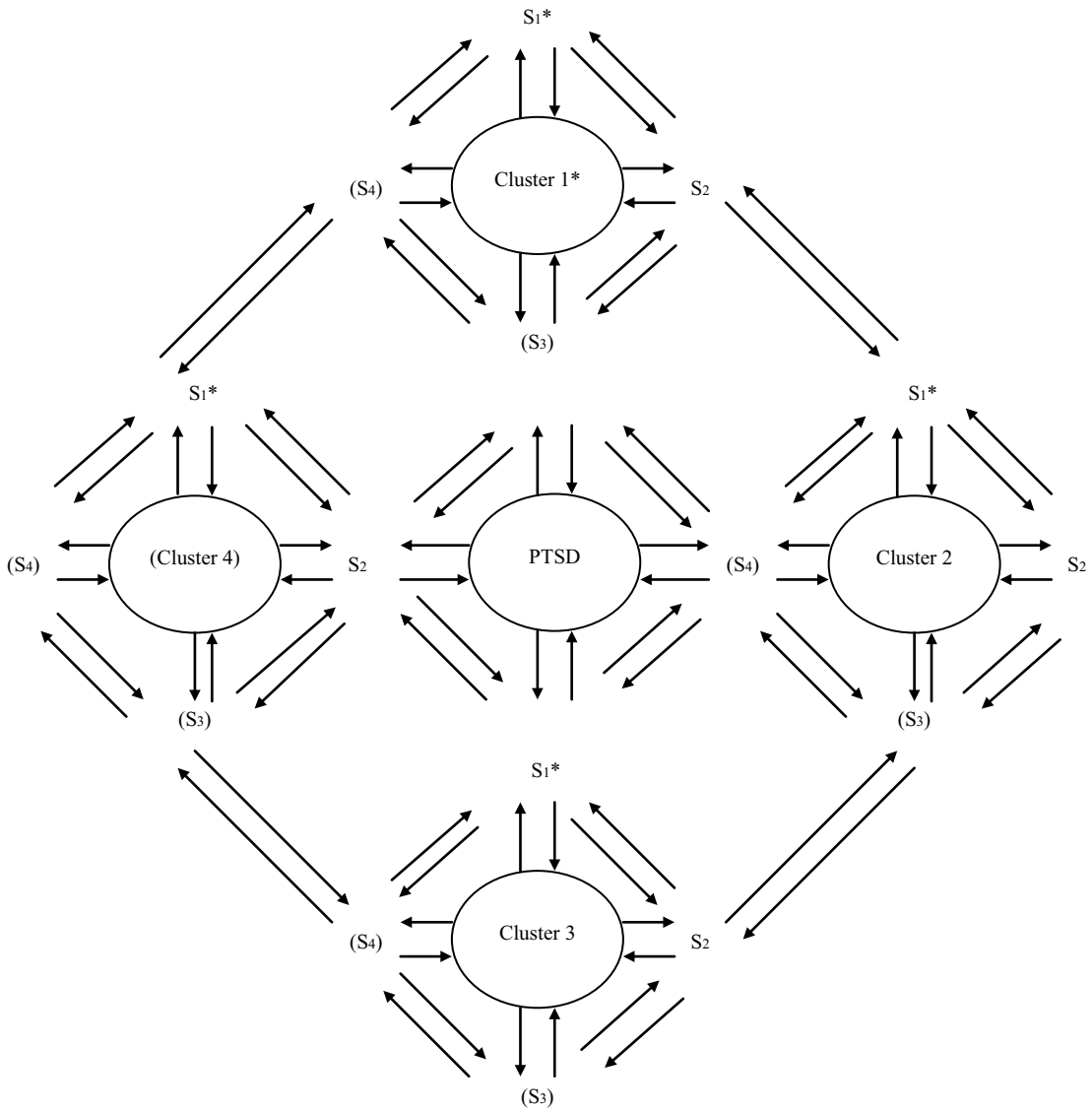


FIGURE 4 | Integrative causal symptom-construct model in mental disorder. The figure depicts the relationship between symptoms and mental disorder (or a symptom cluster of one) as dynamically reciprocal in causation. The mental disorder constitutes an underlying, higher-order level in the patient's mental state symptoms, while the symptoms interact at lower levels of the system, with both the top-down and bottom-up influences dynamically influencing each other in context and over time. Note: the parentheses indicate that PTSD might have only three clusters (as in the DSM-IV), and a cluster might have only two symptoms. Of course, depending on the disorder involved either might have more items (i.e., clusters or symptoms, respectively). Of the clusters in any mental disorder, for their symptoms, it would be beneficial to specify which ones are core/primary. For the model presented in the figure, these could be the first clusters or symptoms that are specified by the asterisks.

To conclude the article, in the next section, I return to considering the nature of symptoms by querying their unconscious, subjective, descriptive, and meaning side compared to their conscious, objective, and reductionist universal causal side. I present a novel model that addresses the question in an integrated manner.

One could ask even whether overarching illness entities could impact symptoms, and that mental disorders could be reducible to symptoms sets, as in the DSM-5. One answer to this conundrum would be to abandon the DSM-5 because of its multiple critics [see Ref. (7, 49)]. For them, the DSM-5 has theoretical, epistemological, and social weaknesses; was the result of a chaotic revision process; does not consider sufficiently the causality related to the listed mental disorders; they are artificial; and so on. However, continued research and revision of its categories could be improving its clinical usage.

Psychiatry needs to address critical questions on the nature of symptoms and mental disorder but, at the same time, balance scienticism and skepticism, or create hybrid models that integrate them and go beyond them. We need pause for thought

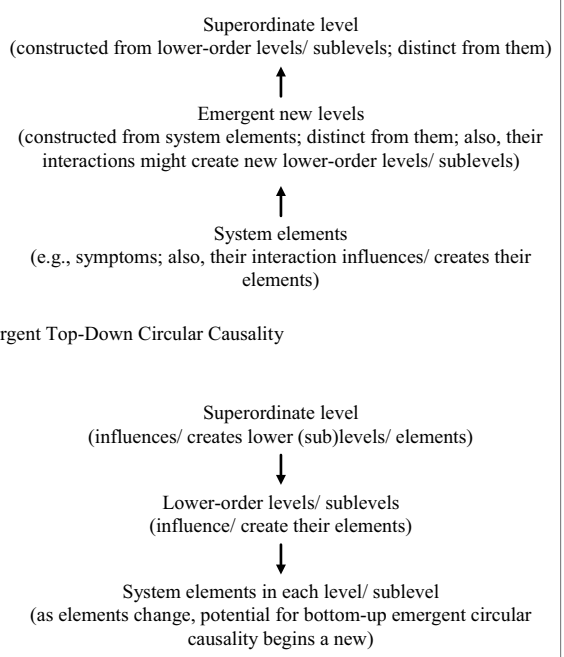


FIGURE 5 | The interaction of top-down and bottom-up emergent circular causality. **(A)** Emergent bottom-up circular causality, **(B)** emergent top-down circular causality.
Note: (1) Configuration/pattern changes possible, too, within and between (sub)levels. (2) Bottom-up and top-down causal processes work together reciprocally in system causality.
The figure illustrates the dynamic interaction of bottom-up and top-down processes both within and across levels in a system, including the possibility of emergence of new symptoms, levels, and sublevels. It also indicates the change of patterning or configuration possible within and between levels in the system dynamics involved. Briefly, as system elements (e.g., symptoms) or levels/sublevels interact, they might influence/create their configuration/ patterning, expression, or even de novo emergence. This process may occur both through movement from lower to higher levels in the level hierarchy involved (bottom-up), or from higher to lower levels (top-down), or reciprocally in both ways. In essence, the figure clarifies that, in system function, bottom-up processes work both within and between levels, as do top-down processes.

in evaluating the relative roles of hermeneutic insight and causal explanation in psychiatry (Verstehen, Erklären, respectively). Whether we accept that symptom meaning/content, or phenomenology can be influenced by hermeneutic insight, "Verstehen" has important consequences. Are symptoms and their meaning/ content only what can be observed, and therefore, reduced to what is measurable, or are there other levels to consider? In this regard, symptom meaning or content might be a higher-order level in the symptom/disorder complex, whether the symptoms are observed or self-reported. Moreover, observed and self-reported symptoms could be tapping different patient realities, and what might these differences mean for symptom meaning/content? For cause, are reductionist, biological views used to explain symptoms/ disorders rather than higher-order mental content or constructs? Can the latter causally influence lower-order (and more easily observed/self-reported) symptoms (Erklären)?

Figure 6 presents a model of symptom expressions that illustrates the difficulty in addressing these types of questions, while proposing a nuanced solution. On the one hand, symptom meaning and causality do not necessarily stand in opposition. For example, at the level of individualization and universalization, symptoms could be unique to the person's history and current mental content, as well as unique in the coalition of forces that had created them. As well, symptoms could reflect universal themes and concerns, and also reflect standard common causal mechanisms.

Ultimately, both individual and universal mental content and symptoms might not be what they appear, either to the person expressing them phenomenologically and subjectively or to the observer using empirical methods, e.g., in observations, interviews, self-report questionnaires, in discerning them. Both subjective and objective understanding of symptoms might approach their reality.

Certainly, a complicating factor in all these regards relates to the play of unconscious processes in symptom creation and expression. This could apply in the sense of (a) classic Freudian repression, (b) automaticity in thought without deliberative reflection or insight, or (c) a lack of awareness of the overall system in which the symptoms are embedded.

Specifically for the area of PTSD and some other related conditions/disorders that are subject to legal dispute (e.g., chronic pain/somatic symptom disorder), the answers to these types of questions are complicated by court considerations. One needs to veer toward the more objective side as much as possible in order to vet possible confounds, such as malingering and unconscious influences on clinical presentation and self-report. **Figure 6** illustrates that intention is very difficult to evaluate and can never be evaluated uniquely by test results or clinical interview. Young (13, 36, 39, 50, 51) has presented work relevant to the question, calling for a scientifically-informed comprehensive impartial approach to assessment in these types of cases.

Ultimately, the network approach to symptom and mental disorder relationship addresses some of the issues raised about individual insight vs. universal explanation in symptomatology because, in this view, how symptoms interact becomes the seat of causal explanation and understanding. Nevertheless, in my hybrid model, one needs to consider top-down psychological construct influences on symptoms as much as their bottom-up interactions, so that their meaning and causation lay in not only symptom networking processes but also in higher-order levels in the symptom structure and the causes associated with them.

Often issues in our field are presented as a dichotomy, or in black and white. For example, for the causes of behavior, too often they are phrased as Nature vs. Nurture. Yet, behavioral causation reflects an interaction of biological, personal (e.g., self, free will belief), and environmental factors (6). Similarly, for the issue of scientism vs. skepticism and how it relates to considering symptoms in terms of individualized meanings or universal (read reductionist) causal mechanisms (Verstehen, Erklären, respectively), the opposition is presented too simply. The question of whether the nature of symptoms are either more unconscious, subjective/phenomenological, and meaningful in content or more conscious, observable, objective, and expressions of universal

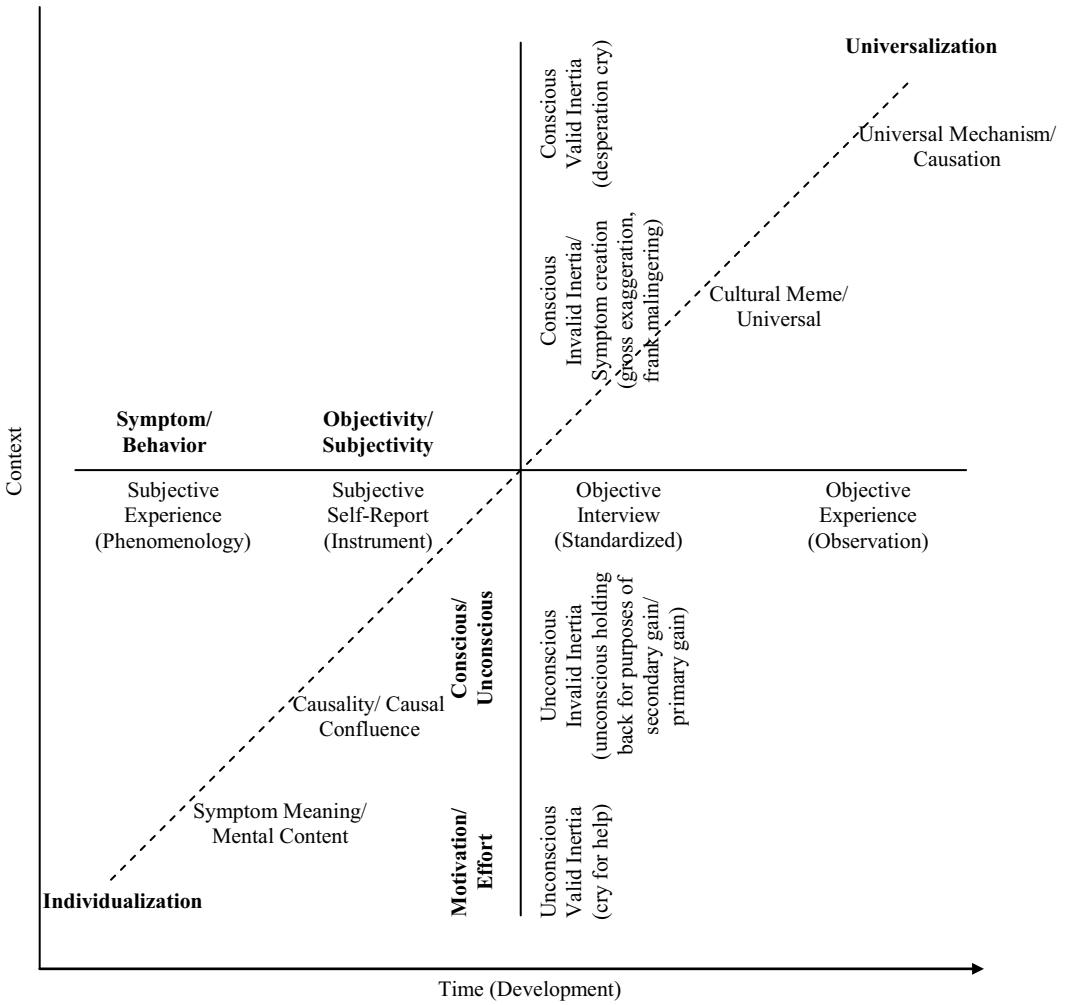


FIGURE 6 | The dimensions of symptom expression in psychiatry. The figure places symptoms along three dimensions that concern: motivation/effort (are symptoms valid?); subjectivity/objectivity (e.g., phenomenological, empirical); and individuality/universality. (a) Symptoms might not be valid, or they could be ones that should not be taken at face value. For example, about the former, there might be conscious malingering for secondary gain that is taking place. About the latter, there might be unconscious influences at work; for example, an unconscious cry for help could take place or a conscious desperation cry. (b) Also, symptoms vary in terms of they are ascertained subjectively or objectively. For example, they might derive from the person's self-report, in particular, or from the evaluator's more controlled efforts to observe/discern them. (c) Finally, symptoms vary in their individuality/universality. At one extreme, each symptom has a personal meaning or value for the patient/client that is unique to her/his history/narrative/experience. Or the causes of the symptoms constitute a unique constellation. On the other hand, they might indicate some cultural universal or meme/mythic narrative. In addition, they might reflect a universal causal or mechanism in behavior to which we are all subject and in constant, invariant ways.

causality might be one that masks a greater underlying complexity in understanding them, such as in the three-dimensional model presented in the figure.

In the article, I have presented an integrated top-down (psychological construct)/bottom-up (symptom network interaction) model of the relationship between symptom and disorder. The same model can be applied to understanding the relationship between mental content and their causes. Because of the multiple levels in systems of behavior, emergent contents can develop at higher-order levels that are not totally reflective of, reducible to, or transcribable from the lower levels, including of the causes involved. Mental contents, such as beliefs, emotions, and desires, can emerge and influence symptoms that might be closer to the lower-order biological or physical substrate, including neurobehaviorally, because of the process of circular emergence and the creation of higher-order levels in behavioral systems that the process allows.

In this regard, one example of higher-order mental state influences on lower-order symptoms is found in the how catastrophic and hopeless thought and related cognitive and emotional processes could cause downward spirals in helplessness and amotivation, and then in the specific symptoms of disorders, such as depression or PTSD. In another example, the personally-exaggerated appraisal of stress that then leads to stress-induced headaches is all too real for many of our patients.

Symptoms and mental disorder co-exist in a system in which all relevant levels need to be recognized and researched. There should be no room for exclusive reductionist or constructionist approaches in understanding them, as both are needed. Reductionism and the search for cause in the most basic biological processes should not be equated with scientism. Nor should seeking emergent phenomena that could influence behavior be treated with skepticism. Behavioral, symptom, and mental content states exist coactively with their causes, and science should examine the relations among all these levels with the clarity that patients deserve.

The present article has presented, hopefully, refined thinking in the area of mental disorder. Further effort along these lines might examine a possible systems model of the definition of mental disorder, one that includes levels for – symptoms, clusters, higher-order mental content, mental disorder, and related concepts such as disability. Similarly, treatment can be conceived systemically, e.g., in terms of cascades that might result from effective treatment shifting the patient into the region of health attractors [e.g., Ref. (27)]. Network concepts can be embedded in systems models and, therefore, the two types of models conceptualized together, and even hybridly, can provide a powerful language for grasping the nature of symptoms, mental disorder, causality, and cure (or treatment).

ACKNOWLEDGMENTS

The reviewers made excellent points and recommendations that greatly improved the manuscript. Many thanks. Grants have been involved in supporting this article.

REFERENCES

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. 5th ed. Washington, DC: American Psychiatric Association (2013).
2. McNally RJ, Robinaugh DJ, Wu GWY, Wang L, Deserno MK, Borsboom D. Mental disorders as causal systems: a network approach to posttraumatic stress disorder. *Clin Psychol Sci* (2014):1–14. doi:10.1177/2167702614553230
3. Young G. *Development and Causality: Neo-Piagetian Perspectives*. New York, NY: Springer Science + Business Media (2011).
4. Markus KA, Borsboom D. Reflective measurement models, behavior domains, and common causes. *New Ideas Psychol* (2013) **31**:54–64. doi:10.1016/j.newideapsych.2011.02.008
5. Melchert TP. *Biopsychosocial Practice: A Science-Based Framework for Behavioral Health Care*. Washington, DC: American Psychological Association (2015).
6. Young G. *Unifying Causality and Psychology: Being, Brain, and Behavior*. New York, NY: Springer Science + Business Media (2016; in press).
7. Frances A. *Essentials of Psychiatric Diagnosis: Responding to the Challenge of DSM-5*. New York, NY: Guilford Press (2013).
8. Thyer BA. The DSM-5 definition of mental disorder: critique and alternatives. In: Probst B, editor. *Essential Clinical Social Work Series: Critical Thinking in Clinical Assessment and Diagnosis*. Cham: Springer (2015). p. 45–68.
9. Vanheule S. *Diagnosis and the DSM: A Critical Review*. Hampshire: Palgrave MacMillan (2014).
10. Insel TR, Cuthbert BN, Garvey MA, Heinssen RK, Pine DS, Quinn KJ, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry* (2010) **167**:748–51. doi:10.1176/appi.ajp.2010.09091379
11. Insel TR, Lieberman JA. *DSM-5 and RDoC: Shared Interests*. The National Institute of Mental Health (2013). Available from: http://www.nimh.nih.gov/news/science-news/2013/dsm-5-and-rdoc-shared-interests.shtml
12. Blumenthal-Barby JS. Psychiatry's new manual (DSM-5): ethical and conceptual dimensions. *J Med Ethics* (2014) **40**:531–6. doi:10.1136/medethics-2013-101468
13. Young G. *Malingering, Feigning, and Response Bias in Psychiatric/Psychological Injury: Implications for Practice and Court*. Dordrecht: Springer Science + Business Media (2014).
14. Wigman JTW, van Os J, Borsboom D, Wardenaar KJ, Epskamp S, Klippel A, et al. Exploring the underlying structure of mental disorders: cross-diagnostic differences and similarities from network perspective using both a top-down and a bottom-up approach. *Psychol Med* (2015) **45**:2375–87. doi:10.1017/S0033291715000331
15. Bielczyk NZ, Buitelaar JK, Glennon JC, Tiesinga PHE. Circuit to construct mapping: a mathematical tool for assisting the diagnosis and treatment in major depressive disorder. *Front Psychiatry* (2015) **6**:29. doi:10.3389/fpsyt.2015.00029
16. Thelen E, Smith LB. Dynamic systems theories. 6th ed. In: Damon W, Lerner RM, editors. *Handbook of Child Psychology: Vol. 1. Theoretical Models of Human Development*. Hoboken, NJ: Wiley (2006). p. 258–312.
17. Vallacher RR, Van Geert P, Nowak A. The intrinsic dynamics of psychological process. *Curr Dir Psychol Sci* (2015) **24**:58–64. doi:10.1177/0963721414551571
18. Gottman J, Swanson C, Swanson K. A general systems theory of marriage: nonlinear difference equation modeling of marital interaction. *Pers Soc Psychol Rev* (2002) **6**:326–40. doi:10.1207/S15327957PSPR0604_07
19. Bak P. *How Nature Works: The Science of Self-Organized Criticality*. New York, NY: Springer-Verlag (1996).
20. Samuelson LK, Jenkins GW, Spencer JP. Grounding cognitive-level processes in behavior: the view from dynamic systems theory. *Top Cogn Sci* (2015) **7**:191–205. doi:10.1111/tops.12129
21. Hayes AM, Yasinski C, Barnes JB, Bockting CLH. Network destabilization and transition in depression: new methods for studying the dynamics of therapeutic change. *Clin Psychol Rev* (2015) **41**:27–39. doi:10.1016/j.cpr.2015.06.007
22. Pe ML, Kircanski K, Thompson RJ, Bringmann LF, Tuerlinckx F, Mestdagh M, et al. Emotion-network density in major depressive disorder. *Clin Psychol Sci* (2015) **3**:292–300. doi:10.1177/2167702614540645
23. Schiepek G, Eckert H, Aas B, Wallot S, Wallot A. *Integrative Psychotherapy: A Feedback-Driven Dynamic Systems Approach*. Göttingen: Hogrefe Publishing (2015).
24. Butner JE, Gagnon KT, Geuss MN, Lessard DA, Story TN. Utilizing topology to generate and test theories of change. *Psychol Methods* (2015) **20**:1–25. doi:10.1037/a0037802
25. Molenaar PCM, Nesselroade JR. 7th ed. In: Overton WF, Molenaar PC, editors. *Handbook of Child Psychology and Developmental Science, Vol. 1. Theory and Method*. Hoboken, NJ: Wiley (2015). p. 652–82.
26. Rabinovich MI, Simmons AN, Varona P. Dynamical bridge between brain and mind. *Trends Cogn Sci* (2015) **19**:453–61. doi:10.1016/j.tics.2015.06.005
27. Wichers M, Wigman JTW, Myin-Germeys I. Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. *Emot Rev* (2015) **7**:362–7. doi:10.1177/1754073915590623
28. World Health Organization. *Mental Disorders* (2015). Available at: http://www.who.int/mediacentre/factsheets/fs396/en/
29. Clarke DE, Narrow WE, Regier DA, Kuramoto SJ, Kupfer DJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, part I: study design, sampling strategy, implementation, and analytic approaches. *Am J Psychiatry* (2013) **170**:43–58. doi:10.1176/appi.ajp.2012.12070998
30. Rosen GM, Lilienfeld SO. Posttraumatic stress disorder: an empirical evaluation of core assumptions. *Clin Psychol Rev* (2008) **28**:837–68. doi:10.1016/j.cpr.2007.12.002
31. Rosen GM, Frueh BC, Elhai JD, Grubaugh AL, Ford JD. Posttraumatic stress disorder and general stress disorder. In: Rosen GM, Frueh BC, editors. *Clinician's Guide to Posttraumatic Stress Disorder*. Hoboken, NJ: John Wiley & Sons (2010). p. 3–31.

32. Armour C, Tsai J, Durham TA, Charak R, Biehn TL, Elhai JD, et al. Dimensional structure of DSM-5 posttraumatic stress symptoms: support for a hybrid anhedonia and externalizing behaviors model. *J Psychiatr Res* (2015) **61**:106–13. doi:10.1016/j.jpsychires.2014.10.012
33. Pietrzak RH, Tsai J, Armour C, Mota N, Harpaz-Rotem I, Southwick SM. Functional significance of a novel 7-factor model of the DSM-5 PTSD symptoms: results from the national health of resilience in veterans study. *J Affect Disord* (2015) **174**:522–6. doi:10.1016/j.jad.2014.12.007
34. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. Washington, DC: American Psychiatric Association (1994).
35. Wang L, Zhang L, Armour C, Cao C, Qing Y, Zhang J, et al. Assessing the underlying dimensionality of DSM-5 PTSD symptoms in Chinese adolescents surviving the 2008 Wenchuan earthquake. *J Anxiety Disord* (2015) **31**:90–7. doi:10.1016/j.janxdis.2015.02.006
36. Young G. Dimensions and dissociation in PTSD in the DSM-5: towards eight core symptoms. *Psychol Inj Law* (2015) **8**:219–32. doi:10.1007/s12207-015-9231-5
37. King D, Leskin G, King L, Weathers F. Confirmatory factor analysis of the Clinician-Administered PTSD scale: evidence for the dimensionality of posttraumatic stress disorder. *Psychol Assess* (1998) **10**:90–6. doi:10.1037/1040-3590.10.2.90
38. Armour C, Hansen M. Assessing DSM-5 latent subtypes of acute stress disorder dissociative or intrusive? *Psychiatr Res* (2015) **225**:476–83. doi:10.1016/j.psychres.2014.11.063
39. Young G, Wang JXT. PTSD-SUDs comorbidities in the context of psychological injury and law. *Psychol Inj Law* (2015) **8**:233–51. doi:10.1007/s12207-015-9229-z
40. Elhai JD, Biehn TL, Armour C, Klopper JJ, Frueh BC, Palmieri PA. Evidence of a unique PTSD construct represented by PTSD's D1-D3 symptoms. *J Anxiety Disord* (2011) **25**:340–5. doi:10.1016/j.janxdis.2010.10.007
41. Tsai J, Armour C, Southwick SM, Pietrzak RH. Dissociative subtype of DSM-5 posttraumatic stress disorder in U.S. veterans. *J Psychiatr Res* (2015) **66-67**:67–74. doi:10.1016/j.jpsychires.2015.04.017
42. Zelazny K, Simms LJ. Confirmatory factor analyses of DSM-5 posttraumatic stress disorder symptoms in psychiatric samples differing in criteria A status. *J Anxiety Disord* (2015) **34**:15–23. doi:10.1016/j.janxdis.2015.05.009
43. Weathers FW, Litz BT, Herman DS, Huska JA, Keane TM. The PTSD checklist (PCL): reliability, validity, and diagnostic utility. *Paper Presented at the Meeting of the International Society for Traumatic Stress Studies*. San Antonio, TX (1993).
44. Li H, Wang L, Shi Z, Zhang Y, Wu K, Liu P. Diagnostic utility of the PTSD Checklist in detecting PTSD in Chinese earthquake victims. *Psychol Rep* (2010) **107**:733–9. doi:10.2466/03.15.20.PR0.107.6.733-739
45. Young G, Lareau C, Pierre B. One quintillion ways to have PTSD comorbidity: recommendations for the disordered DSM-5. *Psychol Inj Law* (2014) **7**:61–74. doi:10.1007/s12207-014-9186-y
46. Conway ARA, Kovacs K. New and emerging models of human intelligence. *WIREs Cogn Sci* (2015) **6**:419–26. doi:10.1002/wcs.1356
47. Borsboom D, Cramer AOJ. Network analysis: an integrative approach to the structure of psychopathology. *Annu Rev Clin Psychol* (2013) **9**:91–121. doi:10.1146/annurev-clinpsy-050212-185608
48. Lewis MD. Bridging emotion theory and neurobiology through dynamic systems modeling. *Behav Brain Sci* (2005) **28**:169–245. doi:10.1017/S0140525X0500004X
49. Demazeux S, Singy P. *The DSM-5 in Perspective: Philosophical Reflections on the Psychiatric Babel*. Dordrecht, Netherlands: Springer Science + Business Media (2015).
50. Young G. Resource material for ethical psychological assessment of symptom and performance validity, including malingering. *Psychol Inj Law* (2014) **7**:206–35. doi:10.1007/s12207-014-9202-2
51. Young G, Drogin E. Psychological injury and law I: causality, malingering, and PTSD. *Mental Health Law Policy J* (2014) **3**:373–416.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

HYPOTHESIS AND THEORY
published: 24 August 2015
doi: 10.3389/fpsyt.2015.00118

Circadian rhythms and mood disorders: are the phenomena and mechanisms causally related?

*William Bechtel**

Department of Philosophy and Center for Circadian Biology, University of California San Diego, San Diego, CA, USA

This paper reviews some of the compelling evidence of disrupted circadian rhythms in individuals with mood disorders (major depressive disorder, seasonal affective disorder, and bipolar disorder) and that treatments such as bright light, designed to alter circadian rhythms, are effective in treating these disorders. Neurotransmitters in brain regions implicated in mood regulation exhibit circadian rhythms. A mouse model originally employed to identify a circadian gene has proven a potent model for mania. While this evidence is suggestive of an etiological role for altered circadian rhythms in mood disorders, it is compatible with other explanations, including that disrupted circadian rhythms and mood disorders are effects of a common cause and that genes and proteins implicated in both simply have pleiotropic effects. In light of this, the paper advances a proposal as to what evidence would be needed to establish a direct causal link between disruption of circadian rhythms and mood disorders.

Keywords: mood disorders, circadian rhythms, mechanistic explanations, causal relations between mechanisms

OPEN ACCESS

Edited by:
Leon De Bruin, VU University Amsterdam, Netherlands

Reviewed by:
Oksana Sorokina, The University of Edinburgh, UK
Kyle B. Gustafson, École Polytechnique Fédérale de Lausanne, Switzerland
Victor Gijsbers, Universiteit Leiden, Netherlands

***Correspondence:**
William Bechtel, Department of Philosophy, University of California San Diego, 9500 Gilman Drive, San Diego, CA 92093-0119, USA
bechtel@ucsd.edu

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry

Received: *21 May 2015*
Accepted: *07 August 2015*
Published: *24 August 2015*

Citation:
Bechtel W (2015) Circadian rhythms and mood disorders: are the phenomena and mechanisms causally related? Front. Psychiatry 6:118. doi: 10.3389/fpsyt.2015.00118

Introduction

Much biological research over the past 200 years has proceeded by delineating individual phenomena and explaining each distinct phenomenon by characterizing the responsible mechanism (1). Discovering mechanisms involves localizing the phenomenon in a responsible system that is taken to be the mechanism and then decomposing that system structurally to discover its component parts and functionally to discover the operations those parts perform (2–4). To show that the parts and operations suffice for the phenomenon requires recomposing the mechanism by mentally rehearsing the operations or employing computational models. When the organization is non-sequential, and the operations, non-linear, computational modeling is often necessary to establish that the mechanism could generate the phenomenon (5). Overall, this has been a highly productive strategy. Much has been learnt about the mechanisms for various biological phenomena, but it has also succeeded in revealing the limitations of the approach. One important limitation is that the supposedly independent phenomena and mechanisms are not nearly as independent as initially thought.

Understanding the ways mechanisms are connected is turning out to be an important challenge in twenty-first century biology and medicine, in part because such connections afford useful ways of intervening on systems to control particular phenomena (e.g., to treat specific diseases). This paper explores the challenges in establishing that the mechanism advanced to explain the fact that one phenomenon is causally affected by that put forward to explain another. As in the project of advancing a philosophical analysis of mechanistic explanation, my aim is to ground an account of what is involved in establishing causal relations between mechanisms in the practice of scientific

researchers. To do this, I focus on two biological phenomena in which important progress has been made in identifying mechanisms – circadian rhythms and moods – and engage in a detailed review of the research and the causal claims that have been advanced as to how the mechanisms relate to each other[1].

As much of the research on mood has focused not on the moods of healthy individuals, but mood disorders such as major depressive disorder (MDD) and bipolar disorder (BD), and these have been related to disruptions of normal circadian rhythms, I will generally treat the related phenomena as mood disorders and disruptions of circadian rhythms. While making for a somewhat complex discussion, this does not pose a substantive problem since disrupted phenomena often serve to guide development of the understanding of the mechanism underlying the normally occurring phenomenon (6).

Progress in characterizing and understanding circadian rhythms and moods has largely stemmed from treating the two independently. But research in the later decades of the twentieth century also revealed connections between the two phenomena by showing that mood disorders are accompanied by disruptions of circadian rhythms. During the last two decades research on both circadian rhythms and mood disorders has increasingly focused on the molecular components of the two mechanisms. This research has revealed that molecular components of the circadian clock mechanism also play a role in mood disorders. This has raised the possibility of causal links between the mechanisms so that either disruptions in the circadian mechanism might be viewed as a cause of mood disorders, or mood disorders might be the cause of altered circadian rhythms. Of course the causality could also go in both directions, but my focus is on another alternative – that despite involving shared components, the two mechanisms are really independent and that what researchers are observing are effects of a common cause or pleotropic effects of common components.

My goal in this paper is not to argue for a particular stance on whether circadian rhythm disorders cause mood disorders; rather I will show how research has raised and addressed the question of possible connections between the two phenomena in the Section "Establishing a Relation between Circadian Rhythms and Mood Disorders" and between the two mechanisms in the Section "Research Implicating Molecular Components of the Circadian Clock in Mood Disorders." I then explore whether a causal connection is supported by the evidence in the Section "How are the Mechanisms Related?" In particular, I am concerned with what it would take to establish a causal connection between these mechanisms. Focusing only on the relation between phenomena, the standard measure of independence is that each can be dissociated from the other. There is evidence that some features of mood disorders and circadian disruptions are independent, but the same is true of different features that are treated as aspects of just one phenomenon. The separateness or relatedness of the mechanisms is what ultimately will determine whether researchers judge the phenomena to be causally linked. Mechanisms comprise parts performing operations, which explains the focus on determining whether parts of the circadian mechanism are also implicated in mood phenomena. However, just knowing that two mechanisms employ the same type of part, even if it performs the same operation in each, does not causally link the two mechanisms. Two car engines may employ pistons, but that does not causally connect them. What is required is that the same individual part is a component of both mechanisms and that the parts of the first mechanism affect the second mechanism differently as the first mechanism is in different states. Only then can researchers relate activities in one mechanism to activities in another mechanism. In the case of circadian and mood mechanisms, one needs to show that a protein, for example, affects mood differently as a consequence of its role in different circadian states. This is a demanding standard that has not yet been realized. But before raising these skeptical worries about how circadian disruptions and mood disorders relate, I turn first to the evidence that has been invoked in relating them. As noted above, my objective is to show that the question of how to understand the relation between the generation of circadian rhythm disorders and mood disorders is a real concern for science, not purely an abstract philosophical concern. Subsequent sections, thus, provide a review of how claims about the relation between circadian rhythms and mood developed and the current state of attempts to evaluate whether a causal claim can be substantiated.

Establishing a Relation between Circadian Rhythms and Mood Disorders

Reports of daily cycles of leaf folding in plants stem from ancient times and were first shown experimentally not to be a response to light when De Mairan (7) placed plants in a dark cupboard and observed that they continued to fold. Other examples of daily rhythms followed, such as, Wunderlich's (8) demonstration that body temperature in humans oscillates by over 1°C/day. Although a variety of researchers tried to maintain that these were responses to external cues, by the time of the International Symposium on Biological Clocks in 1960, the recognition that these oscillations continue but with periods of only approximately 24 h (thus, *circa* + *dies*) when light, temperature, and other environmental cues are removed had established that these rhythms were generated endogenously. That fact, plus the ability of these oscillations to be entrained by light or other cues (referred to as *Zeitgebers*) and the determination that they were maintained with the same period at different temperatures, has come to characterize the phenomenon of circadian rhythmicity. Although many researchers sought clues as to the nature of the mechanism, little progress was made until the 1990s. Instead, during the period 1960–1990, much circadian research focused on developing more detailed accounts of circadian phenomena, such as, how the phase of circadian oscillations is affected by light pulses of different strengths and durations.

Whereas the focus on circadian rhythms was largely motivated by trying to understand the phenomenon as it is normally manifested (disruptions of circadian rhythms were largely used to

[1]The relation between circadian rhythms and mood disorders is just one example of relations between circadian and other phenomena/mechanisms that have been the focus of intense research in recent years. Two other major examples are the relations between circadian rhythms and basic metabolism and circadian rhythms and cognition. See Venkataraman et al. (47) for a review of recent research addressing all three of these linkages between circadian rhythms and other phenomena.

identify the responsible mechanism), mood, like other psychiatric phenomena, is typically characterized in terms of disorders. The identification of *melancholia* stems from ancient times, with the term *depression* acquiring general currency by the end of the nineteenth century. DSM-I, in 1952, included the category of *depressive reaction*, and DSM-II (1968) included *depressive neurosis*. Mania, and the shifting from manic to depressive states, was also recognized in the DSM-II as manic-depressive psychosis. The former was subsequently labeled *unipolar* and the later *bipolar*. The terms *major depressive disorder* (MDD) and *bipolar disorder* (BD) were introduced in the 1970s and incorporated into DSM-III in 1980.

Suggestions of a link between mood disorders and circadian rhythms developed about the same time as the endogenous nature of circadian rhythms was established. Many of these focused on sleep disruptions in patients with mood disorders, for which there were case reports but no systematic studies until Hinton (9) performed a detailed study of currently depressed and recovered patients. He showed less sleep during each hour of the night as well as greater motility in those currently depressed. Sleep, however, is only partially under circadian control and Hinton's, as well as a number of other studies in the 1960s and 1970s, such as, Taub and Berger's (10) examination of the effects of altered sleep patterns on mood in healthy individuals, faced the problem of how to differentiate the effects of disrupted sleep and disrupted circadian rhythms. By employing a forced desynchrony protocol using a light–dark period longer than the circadian system could adapt to, Boivin et al. (11) were able to establish that although subjective happiness declined over each daily awake period, it clearly also oscillated in accord with underlying circadian rhythms (as measured, for example, by core body temperature).

A number of researchers in the 1970s and 1980s demonstrated correlations between measures of mood and measures of circadian rhythms. For example, Kripke et al. (12) found altered circadian periods in the manic-depressive patients they studied and determined that only those with shortened rhythms responded to lithium treatment. In another example, Souetre et al. (13) measured body temperature, plasma cortisol, norepinephrine, thyrotropin, and melatonin concentrations in depressed, recovered, and controls with no diagnosis of depression and demonstrated that the phase remained normal but the amplitude was significantly diminished in depressed participants but returned to normal after recovery.

A different strategy for establishing the linkage between mood disorders and circadian rhythms focused on various therapies found to be effective for mood disorders. Perhaps the best known is light therapy for seasonal affective disorders (SADs). After establishing that sunlight and bright artificial light (2000 lux) can suppress melatonin levels in humans, Lewy et al. (14) employed bright light therapy during 3 h after awakening on a seasonally manic-depressive patient during winter when his depression was greatest. This considerably reduced his depressive symptoms. Based on studies with additional patients, Rosenthal et al. (15) introduced the category SAD and presented further evidence of temporary reduction in depression with bright light therapy (the effects usually ceased when light treatment was stopped). The researchers went beyond the correlation to propose the *phase shift hypothesis*, which holds that depression results from the delayed phase of circadian rhythms (or, in a few cases, from an advanced phase) and that the therapeutic effect of light on depression resulted from shifting the phase of circadian rhythms earlier (or later in phase advanced patients). They supported this with evidence of advance (or delay) in the phase of melatonin expression in treated patients [(16, 17); for a more recent review, see Ref. (18)].

The evidence reviewed in this section demonstrated a connection between the phenomena of mood disorders and circadian rhythms. Mood changes according to circadian time and correlates with a number of other measures of circadian rhythmicity. Therapies that treat mood disorders, such as bright light at dawn, serve to alter the phase of circadian rhythms. The success of light therapy alleviating depression suggested a causal account, whereby circadian rhythms when altered are a causal factor in depression. Light therapy, on this account, restores normal rhythmicity and relieves depression. This idea was picked up in other theoretical proposals, such as, the *social Zeitgeber theory* (19), which proposed that in individuals vulnerable to depression, social stress events can disrupt circadian rhythms and that this in turn leads to depression. But to show that there is a direct causal connection between the phenomena requires more than evidence that they are correlated or even that the same treatments affect both. Information about the mechanism is required. Otherwise, one cannot discount the possibility that neither phenomenon directly affects the other but both phenomena are the product of other factors that are correlated or causally related.

Research Implicating Molecular Components of the Circadian Clock in Mood Disorders

The initial clues to the mechanism responsible for circadian rhythms in animals were developed in the early 1970s when circadian researchers both linked the central clock mechanism in mammals to a structure in the hypothalamus known as the suprachiasmatic nucleus (SCN) (20, 21) and, in fruit flies, identified a gene, *Period* (*Per*)[2], in which mutations caused short or long period oscillations or rendered the organisms arrhythmic (22). However, little progress was made until it was possible to clone *Per* and measure concentrations of its mRNA and protein at different times of day. This revealed that both *Per* mRNA the PER protein oscillated, with the protein peaking several hours after the mRNA. Hardin et al. (23) proposed on this basis a delayed negative feedback mechanism in which the protein PER would feedback on its own gene, temporarily inhibiting its own synthesis until concentrations decayed, at which point the inhibition would cease and new PER could be synthesized (**Figure 1**).

In the 15 years after 1990, many more clock genes were identified, including many homologs between the genes found in fruit

[2]Conventions for naming genes differ between insects and mammals. For this paper, I will adopt the mammalian convention of italics and capitalizing the first letter in the case of genes and upper case Roman for proteins regardless of species.

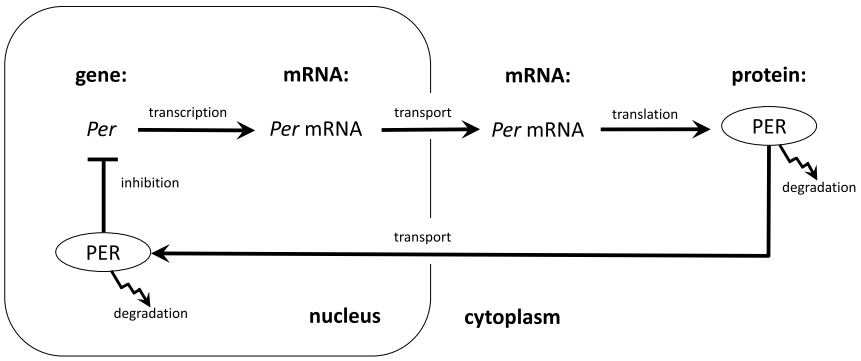


FIGURE 1 | The transcription–translation feedback loop proposed by Hardin et al. (23) to explain circadian rhythms. Transcription of the gene *Per* and translation into protein, PER is followed by the transport of PER back to the nucleus where it inhibits the transcription of its own gene.

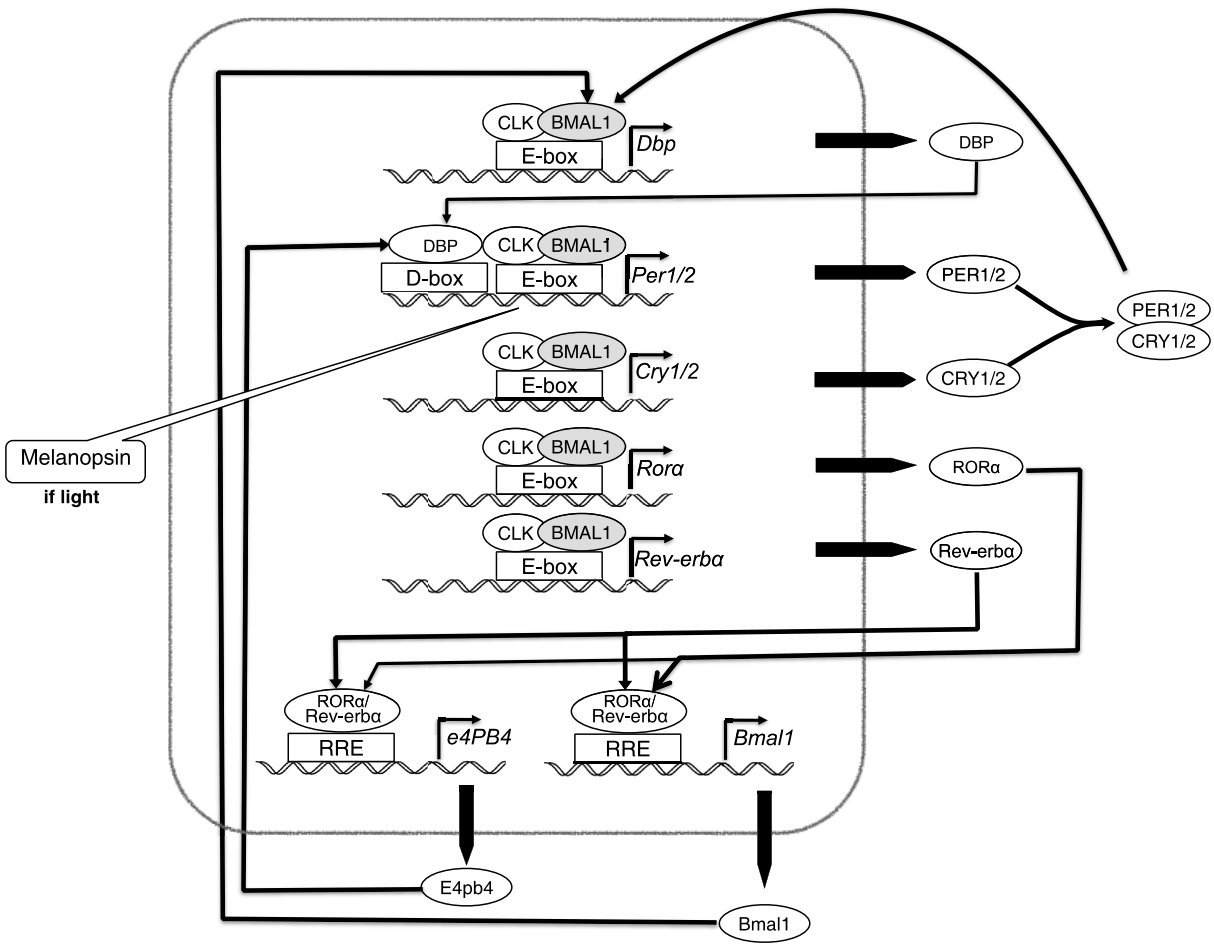


FIGURE 2 | The parts, operations, and organization of the mammalian circadian clock, as understood circa 2005.

flies and mice. Of particular, relevance for the relations between circadian rhythms and mood was the discovery of *Clock* in mice (24), which was shown to bind to the *Per* promoter and to be the target of inhibitory activity by PER. **Figure 2** shows the conception of the circadian mechanism in mammalian cells that had been generated by 2005. Two variant proteins, PER1 and PER2, were now recognized as forming dimers with CRY1 and CRY2, respectively. The inhibition is directed at a dimer between CLOCK (CLK) and BMAL1, which otherwise would bind to the promoter on *Per* and several other genes, activating their transcription. In addition, there is a negative feedback loop involving *E4pb4* and a positive feedback loop involving *Bmal1*.

Much of the research on the mechanisms underlying depression has focused on several monoamine neurotransmitters, especially serotonin, but also norepinephrine and dopamine. These were identified primarily through the fact that many antidepressant drugs increase levels of these monoamines (25). Caspi et al. (26) targeted serotonin as playing a mediating role between stressful life events and depression since individuals with short alleles of serotonin transporter were more prone to experience depression after such events. The different monoamines are associated with particular brain regions that figure centrally in research on depression. The dorsal raphe nuclei are the only source of serotonin in the brain. Dopamine, which is synthesized in decreased amounts in depression, functions in projections from the ventral tegmental area (VTA) to the nucleus accumbens. Norepinephrine is expressed in the locus coeruleus.

One clue to the linkage between circadian rhythms and these neurotransmitters that are altered in depression is that serotonin, norepinephrine, and dopamine all exhibit circadian oscillations in their concentrations (27). Moreover, the connection between the circadian mechanism and synthesis of these monoamines is quite direct: *monoamine oxidase A*, *Maoa*, is a transcriptional target of clock genes *Bmal1* and *Per2* and the protein MAOA serves to terminate dopamine signaling. On the other side, several clock genes have been linked to mood disorders. *Clock*, *Bmal1*, and *Per3* have been implicated in bipolar disease. SNPs of *Per2*, *Npas2*, and *Bmal1* are linked to increased risk for SAD, while there is suggestive evidence of a link between *Cry2* and depression (28). There is also evidence suggestive of a role of mood disorders in affecting circadian rhythms. The SCN has among the densest serotonergic innervation in the brain (all five 5-HT receptor types are employed) and the innervated region of the SCN significantly overlaps areas that receive retinal input and figure in entrainment. This suggests that mood may modulate the ability of the circadian clock to be entrained to local environments, a hypothesis supported by the fact that lesioning the raphe nucleus, thereby eliminating serotonergic innervation of the SCN, alters entrainment. Moreover, applying an agonist of 5-HT receptor generates phase advances (29).

Even stronger evidence indicative of a connection between parts of the circadian mechanism and mood is found in the affects of mutant forms of clock genes on mood. Vitaterna et al.'s discovery that *Clock* is a circadian gene resulted from the generation of a mutant (*Clock*Δ19) that exhibited a long period (27 h) and arrhythmia after several days in darkness. This same mutant has provided a mouse model for mania (30). Both the mutant mice and humans with mania exhibit (1) disrupted circadian rhythms, (2) hyperactivity, and (3) decreased sleep. In addition, the mice exhibit other traits closely resembling those of mania in humans: (4) humans describe feelings of extreme euphoria, while the mice exhibit hyperhedonia and less helplessness, (5) humans engage in increased risk taking, while the mice exhibit reduced anxiety, and (6) while humans exhibit a propensity to drug abuse, the mice show increased preference for cocaine. The mutant mice exhibit increased dopamine in the VTA, rendering neurons there more excitable. As in humans, lithium normalizes manic behavior, and since one effect of lithium is to increase dopamine levels in the VTA, the link appears to be causal.

McClung and her colleagues have investigated *Clock*Δ19 mice to acquire clues into the mechanism underlying human mania. They have found deficits involving the entrainment of low gamma (30–50 Hz) oscillations to delta (1–4 Hz) oscillations in the nucleus accumbens in these mutant mice. In wild-type mice, such entrainment is negatively correlated with amount of exploration in a novel environment. This entrainment is seriously impaired in ClockΔ19 mice, which also exhibit hyperactivity in response to novelty. When treated with lithium, the coupling is restored and the hyperactivity stops (31). Knockdown of *Clock* in the VTA alone results in a manic-like state of less anxiety and hyperactivity but also depressive behavior (32), which seems to fit the fact that manic patients typically also exhibit depressive episodes. Lithium has been shown to lengthen circadian period, likely by inhibiting GSK3β, which phosphorylates PER2 and REV–ERBα, and to produce phase delays as well as affecting the amplitude and period of circadian oscillations (33). It, thus, has the opposite effects on the clock as light treatment, appropriate since it affects mania, not depression.

In an extremely ambitious study, Li et al. (34) provided yet additional evidence pointing to a causal link between circadian genes and mood disorders. They examined gene expression in six brain areas in cadavers of 55 normal controls and 34 patients with MDD post-mortem, analyzing them by circadian time of death. By plotting concentrations of mRNA by the patient's circadian time of death, they generated pseudo-time series data for each gene across the subject pool. In the controls, they identified several hundred genes exhibiting cyclic expression in the dorsolateral prefrontal cortex, amygdala, cerebellum, nucleus accumbens, anterior cingulate cortex, and hippocampus. Many core clock genes were among the genes with the strongest cyclic patterns. These are labeled in yellow on the left in **Figure 3**, in which the *p*-values for those genes with the highest overall significance levels on a measure of oscillation are shown. Red indicates genes whose oscillation is significant ($p < 0.05$) in a given tissue. On the right are the comparable data for the patients with MDD. Comparing the two plots reveals that oscillation reaches statistical significance for many fewer genes in many fewer tissues among the patients. Especially noteworthy is the decrease of rhythmicity in ARNTL (BMAL1) and PER2, two central clock genes, in most brain tissues.

As a measure of the power of the post-mortem gene expression data, Li et al. used the data from 60 randomly selected subjects (both normal controls and patients) to construct an algorithm designed to predict time of death from the pattern of gene expression at death. **Figure 4** shows, for all subjects, the actual time of death in the outer circle and the predicted time of death in the inner circle. Lines connect the corresponding data points and it is clear that the predictions for normal controls align better than for those for patients with MDD. This indicates substantial disruption in the circadian pattern of gene expression in the depressed patients.

Another type of evidence suggestive of a causal connection between the circadian mechanism and that involved in mood disorders is that several of the interventions designed to affect either mood or circadian rhythms also affect the other. One of the most popular drugs used to treat depression, the serotonin reuptake inhibitor (SSRI) fluoxetine, also induces phase advance in the SCN in slices of rat brain in culture. Other SSRIs shorten

A

Symbol	DLPFC	AnCg	HC	AMY	NAcc	CB
ARNTL*	0.0005	0.0005	0.001	0.001	0.0005	0.0005
PER2*	0.001	0.0005	0.0005	0.005	0.008	0.0005
PER3*	0.0005	0.0005	0.0005	0.094	0.0005	0.001
NR1D1*	0.0005	0.0005	0.0005	0.102	0.0005	0.008
DBP*	0.0005	0.0005	0.003	0.066	0.002	0.001
SFPQ	0.0005	0.152	0.013	0.134	0.001	0.029
ITIH5	0.0005	0.021	0.009	0.577	0.007	0.027
LDLR	0.001	0.002	0.004	0.014	0.561	0.204
PER1*	0.0005	0.005	0.008	0.559	0.044	0.071
INSIG1	0.007	0.003	0.001	0.025	0.727	0.183
SLC39A14	0.007	0.0005	0.029	0.067	0.077	0.262
NFIL3*	0.011	0.0005	0.104	0.181	0.03	0.198
SNTB2	0.023	0.038	0.17	0.001	0.013	0.368
PDZRN3	0.002	0.001	0.037	0.333	0.207	0.195
BHLHE40*	0.0005	0.014	0.147	0.42	0.02	0.127
BHLHE41	0.005	0.003	0.997	0.751	0.051	0.003

1 0.8 0.6 0.4 0.2 0

B

MDD:

	DLPFC	ACG	HC	AMY	NACC	CB
ARNTL	0.121	0.072	0.732	0.084	0.005	0.141
PER2	0.015	0.083	0.213	0.342	0.075	0.582
PER3	0.42	0.652	0.432	0.999	0.29	0.531
NR1D1	0.04	0.029	0.285	0.097	0.003	0.111
DBP	0.102	0.236	0.025	0.79	0.121	0.118
SFPQ	0.135	0.124	0.165	0.265	0.047	0.089
ITIH5	0.936	0.47	0.117	0.603	0.15	0.832
LDLR	0.012	0.385	0.315	0.307	0.005	0.028
PER1	0.006	0.21	0.137	0.619	0.124	0.061
INSIG1	0.056	0.534	0.668	0.869	0.318	0.88
SLC39A14	0.641	0.21	0.301	0.393	0.157	0.354
NFIL3*	0.565	0.326	0.633	0.478	0.179	0.617
SNTB2	0.928	0.194	0.123	0.765	0.365	0.293
PDZRN3	0.13	0.003	0.503	0.229	0.075	0.139
BHLHE40	0.19	0.897	0.433	0.963	0.14	0.009
BHLHE41	0.497	0.781	0.754	0.433	0.875	0.246

1 0.8 0.6 0.4 0.2 0

FIGURE 3 | (A) Top oscillating genes in six brain tissues in normal controls in a cadaver study in which time of death was known. Shown in yellow are genes known to figure in the circadian clock. The numbers indicate significance levels on a measure of oscillation and color coding, with color coding differentiating five groups. **(B)** Data for patients with major depressive disorder analyzed in the same manner, making it clear that these genes achieved much lower significance levels in almost all tissues. Reproduced from Li et al. (34), Figures 2 and 4A. Copyright (2013) National Academy of Sciences, USA.

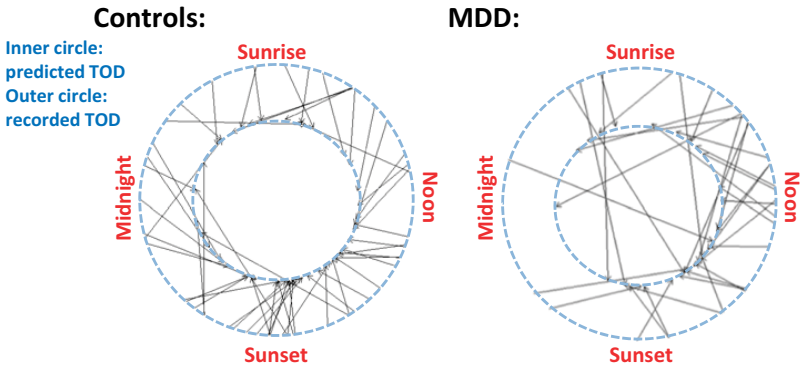


FIGURE 4 | In normal controls, as shown on the left, the pattern of gene expression at death, shown on the inner circle, was a good predictor of time of death, shown on the outer circle. Most of the lines connect nearby points on the two circles. In depressed patients (shown on the right), however, gene expression at death provided a must less reliable predictor of time of death. Reproduced from Li et al. (34), Figure 4C. Copyright (2013) National Academy of Sciences, USA.

the circadian period. Agomelatine, a relatively recent drug that advances the phase in melatonin expression, has proven effective in treating depression (it also, though, functions as an agonist to serotonin 2C receptors so the effects might not just be through altering circadian phase).

In this section, I have discussed the development of mechanistic accounts of both circadian rhythms and mood disorders. What the research has revealed is a host of possible connections between these mechanisms and many have found these highly suggestive that circadian disruption might cause mood disorders, or vice versa. But is such evidence sufficient to establish that the mechanisms are causally linked? I turn to that question in the next section.

How are the Mechanisms Related?

The evidence discussed in the previous sections, claiming both phenomenal and mechanistic connections between circadian rhythms and mood, invites a causal interpretation: disrupted circadian rhythms cause mood disorders (or vice versa). In fact, though, there are four different possibilities to consider:

1. Circadian disruptions cause mood disorders.
2. Mood disorders cause circadian disorders.
3. Causation goes in both directions.
4. Both circadian disruptions and mood disorders are common effects of something else or pleiotropic effects of common components.

Most interpretations adopt the first possibility. Yet, if there is a connection in that direction, there is likely also to be feedback from mood disorders to circadian disorders. However, my concern in this section is with the fourth possibility, which rejects a direct causal connection between circadian rhythms and mood disorders. I will argue that the evidence to date does not allow one to reject possibility.

One way some researchers have tried to support a causal connection is to identify intervening pathways between the two mechanisms. McClung (35) identifies several pathways by which circadian rhythms might regulate mood, of which I discuss three.

There are no direct projections from locus of the central circadian clock, the SCN, to the main loci of mood mechanisms, the dorsal raphe nucleus, the VTA, or the locus coereleus. But McClung shows that there are indirect pathways through a number of hypothalamic nuclei (e.g., the SCN projects to the dorsomedial hypothalamus which then projects to all three areas). These pathways appear to enable circadian oscillations in the SCN to regulate monoamine synthesis in these tissues. These pathways would explain how monoamine levels are altered in mutant mice in which clock genes are mutated or knocked down. McClung identifies a second pathway through the immune system. Alternations to circadian rhythms have effects on the immune system, leading to increased levels of proinflammatory cytokines. Increased levels of proinflammatory cytokines have previously been implicated in depression (36) as well as reduced neurogenesis, neural plasticity, and long-term potentiation. Moreover, the reduction in neurogenesis as well as depressive behaviors can be blocked in environments otherwise inducing stress by applying an inhibitor of nuclear factor-κB (NF-κB). This points to the NF-κB pathway as figuring in generating depression in animals with altered circadian rhythms. The determination that CLOCK itself interacts with NF-κB to activate transcription at NF-κB responsive promoters (37) further supports this as a candidate pathway for linking circadian mechanisms and mood disorders. A third pathway McClung proposed involves glucocorticoids. Concentrations of glucocorticoids increase in stress situations, a condition correlated with mood disorders. A neuronal and hormonal excitatory pathway from the SCN through the paraventricular nucleus (PVN) and the pituitary to the adrenal gland results in the rhythmic synthesis and release of glucocorticoids that then feed back onto the PVN and adrenal glands to maintain stable levels. Two clock proteins figure in regulating glucocorticoid levels. CRY proteins repress glucocorticoid receptors on the PVN and adrenal glands, thereby generating oscillation in the response to glucocorticoids. The receptors are also acetylated by CLOCK, which also decreases their sensitivity to glucocorticoids in the morning and increases it in the evening when acetylation is reversed (38).

Together with the evidence that components of the circadian clock are involved in mood, the evidence of pathways through which the circadian clock could regulate moods makes the case that the circadian clock plays a role in regulating seem plausible. But such evidence alone does not address the question as to whether the mechanisms themselves are actually linked. They may share components, but the roles these components play in each mechanism may be impendent of the role they play in the other. If that were the case, then, even if there are pathways that could connect the two mechanisms, the two mechanisms are not affecting each other – the generation of circadian rhythms is not affecting moods. Each mechanism operates on its own. Establishing that the circadian mechanism is what is affecting moods requires demonstrating that when common components contribute to one phenomenon, they do so in a way that is responsive to their role in the mechanism responsible for the other phenomenon.

Invoking the phenomenon of pleiotropy – the same gene having multiple functions – Landgraf et al. (39) make the case that the mechanisms responsible for circadian rhythms and moods may operate independently while sharing components:

> it is important to point out that, although commonly called 'clock genes', the molecular components of the circadian clock have pleiotropic functions, including many functions that have nothing to do with the clock: manipulating clock genes affects more than just circadian rhythms.

Landgraf et al. marshal their argument by considering several of the kinds of evidence for a causal link from circadian phenomena or mechanisms to mood such as I have presented in previous sections. In response to each, they argue that the evidence is compatible with the circadian and mood mechanisms operating independently while using common components. In fact, given differences in the way the two mechanisms operate, Landgraf et al. suggest there is reason to distinguish, not integrate, the two mechanisms. I will briefly present four of their examples.

One phenomenal linkage between circadian rhythms and mood involves the use of sleep deprivation as a means of transiently ameliorating symptoms of depression in both MDD and BD patients (40, 41). Sleep deprivation has also been found to induce mania in mice (42), which could then be successfully treated with lithium or tamoxifen. Sleep deprivation has been shown as well to have effects on circadian rhythms in mice, hamsters, and humans. A possible explanation for the dual affects of sleep deprivation on both mood and circadian rhythms is that they are the product of one integrated mechanism. But Landgraf et al. note there are also important differences between the phenomena. For example, a brief nap the following day can result in a relapse of depression, but does not have any effect on the circadian system. Sleep is only partially a circadian phenomenon, and it is possible that the effects on depression depend on a non-circadian pathway, perhaps involving cytokines, cortisol, or brain-derived neurotrophic factor.

A second example involves one of the possible pathways that might link circadian disruptions and mood disorders I discussed above: the NF-κB signal transduction pathway. Monje et al. (43) have provided intriguing evidence that the immune system, specifically the NF-κB signal transduction pathway, may be a common cause of circadian disruption and mood disorders, not an intermediary. They appeal to the effects of constant darkness, a known strategy for inducing depression-like behavior in rats. It is known that total darkness results in apoptosis of noradrenergic neurons in the locus coeruleus, serotinergic neurons in the dorsal raphe, and dopaminergic neurons in the VTA. Citing evidence of the role of the immune system in depression, they show that constant darkness results in increased levels of proinflammatory cytokine IL-6 in the hippocampus, leading to increased ERK activation. Manipulation of NF-κB inhibitors indicates that NF-κB played a causal role in mood disorders. The authors also demonstrated a causal effect of NF-κB in decreasing PER2 and increasing BMAL1 levels in the hippocampus. Rather than the effect

of constant darkness on mood being mediated by an effect on the circadian system, the two may be independent consequences mediated by a common immunological pathway.

I turn now to claim about two common components of the circadian and mood mechanisms: *Clock* and *Per2*. Without challenging that the appearance of manic symptoms in the *Clock*Δ19 mouse are a result of higher dopamine levels in the VTA, which are themselves the result of the *Clock* gene mutation, Landgraf et al. note that there is little evidence of endogenous circadian rhythms in the VTA (when projections from the SCN are cut, the VTA exhibits no sustained PER2:LUC rhythms). Rather, when rhythms are found in the VTA, they may be driven by the SCN. In the VTA, where *Clock* has an effect on mood, it may not be performing a circadian function at all. It might instead be an independent contribution of the same genes.

PER2 is one of the proteins most clearly oscillating in brain tissues of normal controls. As we saw in discussing Li et al.'s study, its oscillation is greatly reduced in patients with MDD. Mouse studies have shown that stress, which leads to depression-like behaviors, results in lower amplitude oscillations in PER2 (44). But the effects can be dissociated. PER2 rhythms are restored more rapidly than recovery from depression-like behaviors. The link between PER2 and mood is thought to involve MAOA, which, as discussed above, serves to terminate dopamine signaling. MAOA contains an E-box through which its transcription is made rhythmic by fluctuating PER2 and BMAL1 concentrations (45). However, one *Per2* mutant, $Per2^{Brdm1-/-}$, despite no longer generating rhythmic expression of *Maoa*, still exhibits oscillation in dopamine levels in the VTA and its overall dopamine levels are higher than in the wild type. It exhibits much less depression-like behavior in response to a stressful activity like forced swim. Landgraf et al. suggest that the behavioral changes may be due to decrease in constitutive expression of dopamine, not its oscillation.

These and other examples that Landgraf et al. present illustrate a variety of ways to discount the suggestive links between the phenomena of circadian disruption and mood disorders and between the mechanisms for each. The two phenomena may be independent effects of a common cause[3] and the responses to therapeutic intervention, while similar, may show differences to yet other manipulations. The two mechanisms may involve the same parts without the parts figuring in a common mechanism. The roles in the different mechanisms may be pleiotropic. Moreover, the mood mechanism may not be responsive to the oscillation in dopamine but to how much is expressed. Altogether these results raise doubts about whether the correlations discussed earlier between circadian disruptions and mood disorders are due to a causal linkage between the mechanisms responsible for the two phenomena.

Raising doubts about whether two phenomena or their respective mechanisms are causally connected is very different from showing that they are not. Given the variety of correlations between mood disorders and circadian disruptions, it seems highly plausible that they are causally connected. What the challenges by Landgraf et al. indicate is that different evidence is required to establish a causal link than has been provided by most of the research to date. What would seem to be required is a demonstration that perturbations that alter one phenomenon affect the other in virtue of the way they altered the first phenomenon. This might be most clearly shown in by focusing on common parts of the circadian mechanism and the mood mechanism and the way they function in each mechanism. If the effect the part has on mood depends on how it behaves in the circadian mechanism that would indicate a causal link between the mechanisms. Experimental perturbations would provide the most compelling evidence: if perturbing how a part functions within the circadian clock thereby perturbs how it behaves in the mood mechanism, then one would have strong evidence that the two mechanisms are causally integrated. Procuring such evidence will be experimentally challenging and the evidence available to date does not support such direct causal connections.

Conclusion

Research on both circadian rhythms and mood disorders has been pursued in the quest for mechanistic explanations of each. This requires delineating the respective phenomena, linking the phenomena with a particular mechanism, and decomposing that mechanism into its parts and operations. I have reviewed several major studies that provide evidence that the phenomena of altered circadian rhythms and mood disorders are correlated and of common components in that the respective mechanisms share components. This research strongly suggests that altered circadian rhythms are a causal factor in the generation of mood disorders. Drawing on arguments advanced by Landgraf et al., however, I have argued that the evidence advanced so far does not yet justify such a claim. The phenomena may be due to common causes and the components of the mechanisms may function independently in each one. My goal, however, is not simply to make the point that correlation does not establish causation but rather to point to what would provide stronger support for the claim that circadian disruptions cause mood disorders. What is required is to show that it is as a contributor to the circadian mechanism that a component affects moods. One way this can be done is to establish that as its state in the circadian mechanism changes, the contribution of the component to the generation of moods and mood disorders changes.

The general issue of how mechanisms hypothesized to explain different phenomena are connected to one another is becoming increasingly important in biology and medicine. The investigation of proteins or the genes that code for them has often revealed components that figure in two or more cellular functions, sometimes in the same tissue. The question, as I have tried to argue here, is then whether the common component actually provides a causal linkage between the mechanisms. One factor accelerating the need to address this question is that researchers are moving beyond classical techniques that typically only identify a handful of very important components to a mechanism, such as *Per*, *Cry*, *Clock*, and *Bmal1*, in the case of circadian rhythms. Newer techniques reveal a plethora of components. For example, using small interfering RNAs to screen the whole

[3]If one opens the investigation into common causes, one finds many possible candidates. In footnote 1, I alluded to the growing evidence of multiple points of connection between circadian mechanisms and metabolic processes. Metabolic processes are one for common causes between the circadian mechanisms and mood mechanisms.

genome, Zhang et al. (46) identified an additional 200 genes beyond those thought to constitute the core clock which have affects on the amplitude or period of circadian rhythms. Many of these genes were previously characterized in terms of their roles in other cellular functions and it becomes relevant to know whether they are operating within their role in these other cellular functions in affecting circadian rhythms. If they do, then manipulating these other phenomena may be a viable strategy for altering circadian rhythms, or vice versa.

My concern has been to present issues of this sort as arising within ongoing science. Accordingly, I have examined in detail how the issue has arisen in the case of circadian rhythm research and mood. But the issue is a general one. Finding relations between phenomena and common parts within the mechanisms proposed to explain them can be very suggestive of a causal relation. Yet, until one has shown that a component has its effect on one mechanism in virtue of its role in the other, one has not established a causal connection between the phenomena.

References

1. Bechtel W, Richardson RC. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. 1993 ed. Cambridge, MA: MIT Press, Princeton University Press (1993/2010).
2. Bechtel W, Abrahamsen A. Explanation: a mechanist alternative. *Stud Hist Philos Biol Biomed Sci* (2005) **36**:421–41. doi:10.1016/j.shpsc.2005.03.010
3. Craver CF, Darden L. *In Search of Mechanisms: Discoveries Across the Life Sciences*. Chicago, IL: University of Chicago Press (2013).
4. Machamer P, Darden L, Craver CF. Thinking about mechanisms. *Philos Sci* (2000) **67**:1–25. doi:10.1086/392759
5. Bechtel W, Abrahamsen A. Dynamic mechanistic explanation: computational modeling of circadian rhythms as an exemplar for cognitive science. *Stud Hist Philos Sci* (2010) **41**:321–33. doi:10.1016/j.shpsa.2010.07.003
6. Simon HA. *The Sciences of the Artificial*. 2nd ed. Cambridge, MA: MIT Press (1980).
7. De Mairan J. Observation botanique. *Histoire de l'Académie royale des sciences* (1729). p. 35.
8. Wunderlich KRA. *Das Verhalten der Eigenwärme in Krankheiten*. Leipzig: Otto Wigard (1868).
9. Hinton JM. Patterns of insomnia in depressive states. *J Neurol Neurosurg Psychiatry* (1963) **26**:184–9. doi:10.1136/jnnp.26.2.184
10. Taub JM, Berger RJ. Performance and mood following variations in the length and timing of sleep. *Psychophysiology* (1973) **10**:559–70. doi:10.1111/j.1469-8986.1973.tb00805.x
11. Boivin DB, Czeisler CA, Dijk DJ, Duffy JF, Folkard S, Minors DS, et al. Complex interaction of the sleep-wake cycle and circadian phase modulates mood in healthy subjects. *Arch Gen Psychiatry* (1997) **54**:145–52. doi:10.1001/archpsyc.1997.01830140055010
12. Kripke DF, Mullaney DJ, Atkinson M, Wolf S. Circadian rhythm disorders in manic-depressives. *Biol Psychiatry* (1978) **13**:335–51.
13. Souetre E, Salvati E, Belugou JL, Pringuey D, Candito M, Krebs B, et al. Circadian rhythms in depression and recovery: evidence for blunted amplitude as the main chronobiological abnormality. *Psychiatry Res* (1989) **28**:263–78. doi:10.1016/0165-1781(89)90207-2
14. Lewy AJ, Kern HA, Rosenthal NE, Wehr TA. Bright artificial light treatment of a manic-depressive patient with a seasonal mood cycle. *Am J Psychiatry* (1982) **139**:1496–8. doi:10.1176/ajp.139.11.1496
15. Rosenthal NE, Sack DA, Gillin JC, Lewy AJ, Goodwin FK, Davenport Y, et al. Seasonal affective disorder. A description of the syndrome and preliminary findings with light therapy. *Arch Gen Psychiatry* (1984) **41**:72–80. doi:10.1001/archpsyc.1984.01790120076010
16. Lewy AJ, Sack RL, Singer CM, White DM. The phase shift hypothesis for bright light's therapeutic mechanism of action: theoretical considerations and experimental evidence. *Psychopharmacol Bull* (1987) **23**:349–53.
17. Lewy AJ, Sack RL, Singer CM, White DM, Hoban TM. Winter depression and the phase-shift hypothesis for bright light's therapeutic effects: history, theory, and experimental evidence. *J Biol Rhythms* (1988) **3**:121–34. doi:10.1177/074873048800300203
18. Lewy AJ, Emens JS, Songer JB, Sims N, Laurie AL, Fiala SC, et al. Winter depression: integrating mood, circadian rhythms, and the sleep/wake and light/dark cycles into a bio-psycho-social-environmental model. *Sleep Med Clin* (2009) **4**:285–99. doi:10.1016/j.jsmc.2009.02.003
19. Ehlers CL, Frank E, Kupfer DJ. Social zeitgebers and biological rhythms: a unified approach to understanding the etiology of depression. *Arch Gen Psychiatry* (1988) **45**:948–52. doi:10.1001/archpsyc.1988.01800340076012
20. Moore RY. Retinohypothalamic projection in mammals: a comparative study. *Brain Res* (1973) **49**:403–9. doi:10.1016/0006-8993(73)90431-9
21. Moore RY, Eichler VB. Loss of a circadian adrenal corticosterone rhythm following suprachiasmatic lesions in the rat. *Brain Res* (1972) **42**:201–6. doi:10.1016/0006-8993(72)90054-6
22. Konopka RJ, Benzer S. Clock mutants of *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* (1971) **89**:2112–6. doi:10.1073/pnas.68.9.2112
23. Hardin PE, Hall JC, Rosbash M. Feedback of the *Drosophila period* gene product on circadian cycling of its messenger RNA levels. *Nature* (1990) **343**:536–40. doi:10.1038/343536a0
24. Vitaterna MH, King DP, Chang A-M, Kornhauser JM, Lowrey PL, McDonald JD, et al. Mutagenesis and mapping of a mouse gene, *Clock*, essential for circadian behavior. *Science* (1994) **264**:719–25. doi:10.1126/science.8171325
25. Nutt DJ. Relationship of neurotransmitters to the symptoms of major depressive disorder. *J Clin Psychiatry* (2008) **69**(Suppl E1):4–7.
26. Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, et al. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* (2003) **301**:386–9. doi:10.1126/science.1083968
27. Hampp G, Ripperger JA, Houben T, Schmutz I, Blex C, Perreau-Lenz S, et al. Regulation of monoamine oxidase A by circadian-clock components implies clock influence on mood. *Curr Biol* (2008) **18**:678–83. doi:10.1016/j.cub.2008.04.012
28. Albrecht U. Circadian clocks and mood-related behaviors. In: Kramer A, Merrow M, editors. *Circadian Clocks*. (Vol. 217), Berlin: Springer (2013). p. 227–39.
29. Sollars PJ, Pickard GE, Sprouse JS. Serotonin and the regulation of mammalian circadian rhythms. In: Squire LR, editor. *Encyclopedia of Neuroscience*. Oxford: Academic Press (2009). p. 723–30.
30. Roybal K, Theobold D, Graham A, DiNieri JA, Russo SJ, Krishnan V, et al. Mania-like behavior induced by disruption of CLOCK. *Proc Natil Acad Sci U S A* (2007) **104**:6406–11. doi:10.1073/pnas.0609625104
31. McClung CA. Circadian rhythms and mood regulation: insights from preclinical models. *Eur Neuropsychopharmacol* (2011) **21**(Suppl 4):S683–93. doi:10.1016/j.euroneuro.2011.07.008
32. Mukherjee S, Coque L, Cao J-L, Kumar J, Chakravarty S, Asaithamby A, et al. Knockdown of *Clock* in the ventral tegmental area through RNA interference results in a mixed state of mania and depression-like behavior. *Biol Psychiatry* (2010) **68**:503–11. doi:10.1016/j.biopsych.2010.04.031
33. Li JZ, Lu WQ, Beesley S, Loudon AS, Meng QJ. Lithium impacts on the amplitude and period of the molecular circadian clockwork. *PLoS One* (2012) **7**:e33292. doi:10.1371/journal.pone.0033292
34. Li JZ, Bunney BG, Meng F, Hagenauer MH, Walsh DM, Vawter MP, et al. Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. *Proc Natl Acad Sci U S A* (2013) **110**:9950–5. doi:10.1073/pnas.1305814110
35. McClung CA. How might circadian rhythms control mood? Let me count the ways. *Biol Psychiatry* (2013) **74**:242–9. doi:10.1016/j.biopsych.2013.02.019
36. Miller AH, Maletic V, Raison CL. Inflammation and its discontents: the role of cytokines in the pathophysiology of major depression. *Biol Psychiatry* (2009) **65**:732–41. doi:10.1016/j.biopsych.2008.11.029
37. Spengler ML, Kuropatwinski KK, Comas M, Gasparian AV, Fedtsova N, Gleiberman AS, et al. Core circadian protein CLOCK is a positive regulator of NF-κB-mediated transcription. *Proc Natil Acad Sci U S A* (2012) **109**:E2457–65. doi:10.1073/pnas.1206274109
38. Charmandari E, Chrousos GP, Lambrou GI, Pavlaki A, Koide H, Ng SSM, et al. Peripheral CLOCK regulates target-tissue glucocorticoid receptor transcriptional activity in a circadian fashion in man. *PLoS One* (2011) **6**:e25612. doi:10.1371/journal.pone.0025612

39. Landgraf D, McCarthy MJ, Welsh DK. The role of the circadian clock in animal models of mood disorders. *Behav Neurosci* (2014) **128**:344–59. doi:10.1037/a0036029
40. Giedke H, Schwarzler F. Therapeutic use of sleep deprivation in depression. *Sleep Med Rev* (2002) **6**:361–77. doi:10.1016/S1087-0792(02)90235-2
41. Selvi Y, Gulec M, Agargun MY, Besiroglu L. Mood changes after sleep deprivation in morningness-eveningness chronotypes in healthy individuals. *J Sleep Res* (2007) **16**:241–4. doi:10.1111/j.1365-2869.2007.00596.x
42. Armani F, Andersen ML, Andreatini R, Frussa R, Tufik S, Galduroz JCF. Successful combined therapy with tamoxifen and lithium in a paradoxical sleep deprivation-induced mania model. *CNS Neurosci Ther* (2012) **18**:119–25. doi:10.1111/j.1755-5949.2010.00224.x
43. Monje FJ, Cabatic M, Divisch I, Kim E-J, Herkner KR, Binder BR, et al. Constant darkness induces IL-6-dependent depression-like behavior through the NF-κB signaling pathway. *J Neurosci* (2011) **31**:9075–83. doi:10.1523/JNEUROSCI.1537-11.2011
44. Jiang WG, Li SX, Zhou SJ, Sun Y, Shi J, Lu L. Chronic unpredictable stress induces a reversible change of PER2 rhythm in the suprachiasmatic nucleus. *Brain Res* (2011) **1399**:25–32. doi:10.1016/j.brainres.2011.05.001
45. Hampp G, Albrecht U. The circadian clock and mood-related behavior. *Commun Integr Biol* (2008) **1**:1–3. doi:10.4161/cib.1.1.6286
46. Zhang EE, Liu AC, Hirota T, Miraglia LJ, Welch G, Pongsawakul PY, et al. A genome-wide RNAi screen for modifiers of the circadian clock in human cells. *Cell* (2009) **139**:199–210. doi:10.1016/j.cell.2009.08.031
47. Venkataraman A, Ballance H, Hogenesch JB. The role of the circadian system in homeostasis. In: Walhout AJM, Vidal M, Dekker J, editors. *Handbook of Systems Biology: Concepts and Insights*. Amsterdam: Elsevier (2013). p. 407–26.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in PSYCHIATRY

REVIEW ARTICLE
published: 26 February 2015
doi: 10.3389/fpsyt.2015.00029

Circuit to construct mapping: a mathematical tool for assisting the diagnosis and treatment in major depressive disorder

***Natalia Z. Bielczyk*[1,2]*, *Jan K. Buitelaar*[1,2], *Jeffrey C. Glennon*[1,2] and *Paul H. E. Tiesinga*[1,3]**

[1] *Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands*
[2] *Department of Cognitive Neuroscience, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands*
[3] *Department of Neuroinformatics, Radboud University Nijmegen, Nijmegen, Netherlands*

Edited by:
Annemarie Kalis, Utrecht University, Netherlands

Reviewed by:
Satyaprakash Nayak, Pfizer Inc., USA
Femke L. Truijens, University of Ghent, Belgium

***Correspondence:**
Natalia Z. Bielczyk, Laboratory of Translational Neuroscience, Department of Cognitive Neuroscience, Radboud University Nijmegen Medical Centre, Geert Groteplein 12, Route 126, Nijmegen 6525 GA, Netherlands
e-mail: natalia.bielczyk@radboudumc.nl

Major depressive disorder (MDD) is a serious condition with a lifetime prevalence exceeding 16% worldwide. MDD is a heterogeneous disorder that involves multiple behavioral symptoms on the one hand and multiple neuronal circuits on the other hand. In this review, we integrate the literature on cognitive and physiological biomarkers of MDD with the insights derived from mathematical models of brain networks, especially models that can be used for fMRI datasets. We refer to the recent NIH research domain criteria initiative, in which a concept of "constructs" as functional units of mental disorders is introduced. Constructs are biomarkers present at multiple levels of brain functioning – cognition, genetics, brain anatomy, and neurophysiology. In this review, we propose a new approach which we called circuit to construct mapping (CCM), which aims to characterize causal relations between the underlying network dynamics (as the cause) and the constructs referring to the clinical symptoms of MDD (as the effect). CCM involves extracting diagnostic categories from behavioral data, linking circuits that are causal to these categories with use of clinical neuroimaging data, and modeling the dynamics of the emerging circuits with attractor dynamics in order to provide new, neuroimaging-related biomarkers for MDD. The CCM approach optimizes the clinical diagnosis and patient stratification. It also addresses the recent demand for linking circuits to behavior, and provides a new insight into clinical treatment by investigating the dynamics of neuronal circuits underneath cognitive dimensions of MDD. CCM can serve as a new regime toward personalized medicine, assisting the diagnosis and treatment of MDD.

Keywords: major depressive disorder, modeling, circuit, diagnosis, research domain criteria project, dynamical systems

INTRODUCTION

MAJOR DEPRESSIVE DISORDER

Major depressive disorder (MDD), also known as unipolar depression, has a lifetime prevalence that exceeds 16% in the US (1), and is expected to increase their share in the global disease burden from 4.3% in 2004 to 6.2% by 2030 (2). Treating MDD is costly. In 2010, the total cost of MDD in the EU was estimated to be €798 billion, of which 60% was direct costs and 40% due to lost productivity (3). Currently, there is a rich variety of competing biomarker sets, each suggesting different MDD etiology. However, it is unclear how these relate to the current diagnostic criteria. This heterogeneity of biomarkers, behavioral symptoms, and circuit changes in MDD requires the use of multimodal and multidisciplinary approaches together with mathematical modeling in order to integrate these findings into diagnostic and intervention tools useful in clinical practice.

So far, the search for candidate genes underlying MDD has not yielded a single responsible gene. Instead, genetic models of MDD propose that a large number of genes is involved (4), with a small contribution of each of them to MDD phenotype. Furthermore, these models suggest that epigenetic regulation may underlie critical gene-environment effects in MDD (5). Epidemiological studies have revealed that genetic factors may account for 40–50% of the risk of developing the disorder (6). Since the definition of an endophenotype involves heritability (7) and can only be used in a family sensitive design (8), it leads to a conclusion that only particular diagnostic categories in MDD can be interpreted as endophenotypes. Therefore, instead of talking about endophenotypes in MDD, we refer to NIH research domain criteria (RDoC) project approach (9) and to its central concept of a *construct* as a basic dimension of brain functioning (without a requirement of heritability). While defining constructs, RDoC initiative refers to various units of analysis, from genes to neural circuits and behavior.

In section "Etiology of MDD", we review the current state of knowledge about MDD etiology across multiple construct domains, from behavioral through physiological down to neuronal level. Furthermore, we propose a new paradigm to aid in the diagnosis of MDD and its clinical management which includes dynamical models of the underlying circuitry and mapping the

activity of these circuits onto cognitive constructs diagnostic for MDD. This circuit to construct mapping (CCM) approach can facilitate a personalized approach to MDD and thereby improve the quality of life for MDD patients.

CAUSALITY

Mapping the activity of underlying circuits onto cognitive constructs diagnostic for MDD involves assumption that we can point to causal relations between these two domains. In this review, we focus on the altered dynamics of neuronal circuits as the cause of disrupted behavior. But how can one determine causality? There are two definitions of causality, and both of which are often used in research. First definition by Lewis (10) describes causality in the language of *counterfactuals*: we may define a cause to be an object followed by another, where, if the first object had not been, the second never had existed. On the basis of this definition, in 1986, Holland formulated the "no causation without manipulation" rule (11) which became the prevailing principle in causal research for another two decades. Today, Woodward's view at causality through structural equations comes popular (12). Assuming that we have an endogenous variable Y, produced from variables $X_1, X_2, \ldots, X_n$, Woodard's approach involves expressing certain basic counterfactuals in the following form: *If it were the case that* $X_1 = x_1$, $X_2 = x_2, \ldots, X_n = x_n$, *then it would be the case that* $Y = f(x_1, \ldots, x_n)$.

However, this is not the only view on causality. Judea Pearl builds in the counterfactual approach and writes in his recent essays (13): "the essential ingredient of causation is responsiveness, namely, the capacity of some variables to respond to variations in other variables, regardless of how those variations came about." This is an objection to the idea that the establishment of causation necessarily requires manipulation; rather, it is sufficient to observe the system and its natural course. However, the inference of causality on the basis of observational data is not easy, and Pearl developed a comprehensive theory of how to establish causation by means of probabilistic models.

This latter view of causality is beneficial to causal research in psychiatry; because, we are not always equipped with tools to manipulate all the candidate causes in our system. For instance, if we are interested in the causal effect of the insular cortex on emotional states in patients with MDD and we aim to apply the counterfactual approach in order to test this hypothesis, we should shut down the activity of the isolated insula and register the observed change in regulation of emotional states in our cohort. However, since the insula does not lay on the surface of the cortex, it is very hard to non-invasively perturb its activity alone; since, so far the remote control of deep brain activity is not available in humans. Therefore, in clinical trials the second definition of causality is typically applied: one compares a population of subjects with and without overactivation in the insular cortex, and tries to find systematic differences between these two groups in terms of emotional states. If the effect size is large enough for the groups of a given amount of patients, the causal effect is determined. In the further sections, we will discuss causality in Pearl's sense, meaning "observation" and "statistical power" rather than "intervention" and "counterfactuals."

ETIOLOGY OF MDD

CONSTRUCTS IN MDD

Causality in case of MDD (and other cognitive disorders) is a complex research problem because the disorder can be described across various domains, from neurophysiology, through neuronal networks, to behavior. Although a causal explanation in MDD can search for relationships between any pair of constructs, from the psychiatric point of view links in which behavioral constructs are the effect are especially valuable.

Figure 1 presents the variety of constructs across multiple levels of description in a process of a typical treatment in MDD, with arrows denoting causal relations between them. Firstly, one can distinguish five classes of drugs (**Figure 1A**) on the basis of monoamine receptors that they target (**Figure 1B**). A patient diagnosed with MDD is typically prescribed with one or, rarely, with a combination of these drug types. Functional MRI studies reveal that these drugs affect different, but overlapping circuits (**Figure 1C**). For any given construct, the underlying neuronal circuitry, modulated by interplay between the neural substrates within, reaches a stable activity pattern – which is pictured with the ball metaphor (**Figure 1D**). The network specific activation pattern, as we believe, modulates the particular cognitive construct (**Figure 1E**). The behavior of the patient is subject to repetitive diagnoses which, possibly, can lead to prescription of new, more accurate drugs which closes the circle. In our understanding, the mechanism underlying MDD is a superposition of multiple circuits, each of them having a causal effect on one of the cognitive constructs present in MDD. Therefore, in our considerations on modeling MDD, we are interested in the causal effect between neuronal circuits (as the cause, C) and behavioral constructs (as the effect, D).

We briefly review the aforementioned levels of the description in the following sections. Although the proposed CCM approach includes only mapping from neuronal circuitries straight to the cognitive domain, the physiology underlying MDD is also worth mentioning; because, the most popular (but not necessarily the most effective) treatments derive from the monoamine theory of MDD and target neuromodulatory receptors in the brain rather than particular circuits.

COGNITIVE CONSTRUCTS

Major depression was originally defined in terms of behavior; therefore, cognitive constructs present in MDD seem to be the right starting point to give full characteristics of this disorder. In DSM-5, diagnostic criteria for MDD are as follows: if the subject is diagnosed with MDD if at least five out of nine diagnostic traits are present (**Figure 1E**), at least one of them being anhedonia or low mood.

Current diagnostic practice for MDD is difficult. First, both DSM-5 and ICD-10 diagnostic criteria allow for a broad range of behavioral profiles, all diagnosed with the same clinical condition (14, 15). Second, the diagnostic criteria are open to different interpretations, change over time and are therefore less objective and require review by trained clinicians. For example, independent symptoms of dysthymia (present in DSM-4 as a self-standing disorder) were recently classified as chronic MDD in DSM-5, because since DSM-4 was released there was not enough evidence that

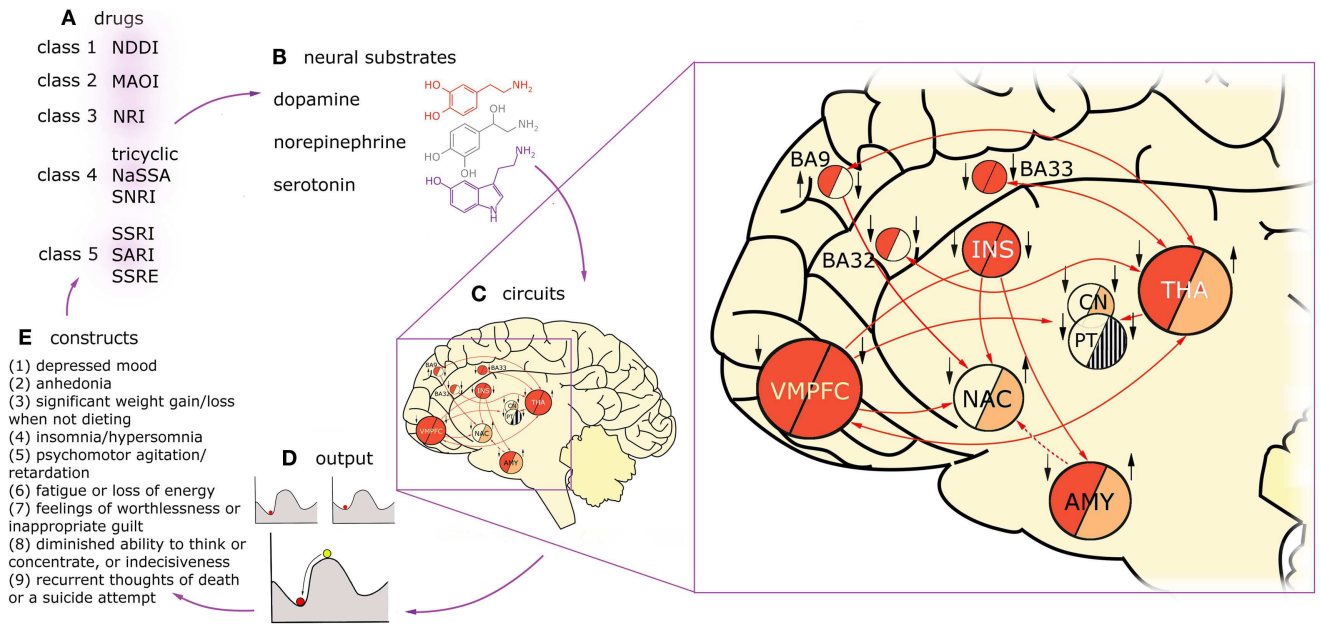


FIGURE 1 | Diagnosis, treatment, and brain dynamics in MDD. One can distinguish five classes of drugs **(A)** on the basis of monoamine receptors that they target **(B)**. Functional MRI studies reveal that these drugs affect different, but overlapping circuits **(C)**. Here, two exemplary circuits are presented on the basis of imaging studies on two constructs present in MDD: low mood (or negative affect) and anhedonia (lack of positive affect). The connectivity in the circuits is presented with arrows, solid lines for glutamatergic, and dashed lines for GABAergic projections. Findings on regions activated in negative affect in depressed patients are summarized in the "negative affect" circuit, left half-circles. The subjects were triggered to fall into low mood by presenting them with scenes of negative emotional valence. Findings on regions up- and down-regulated in anhedonic MDD patients during presentation of scenes with positive emotional valence are summarized in the "anhedonia" circuit, right half-circles. Deep red color depicts overactivation in a given region during the task in respect to healthy controls, light red – hypoactivation, stripes – counteracting evidence in the literature, background color – no data. Additionally, treatment effects of fluoxetine in the nodes of this network are indicated with black arrows (up: up-regulation in respect to non-medicated patients; down: down-regulation). Influence of the drug was assessed on the basis of imaging studies that were using experimental tasks focusing on emotion processing (as aforementioned tasks involving presentation with scenes of emotional valence). On the left hand side of each region, influence of fluoxetine treatment in negative affect regime in MDD is indicated, on the right hand side: the same for positive affect regime. On one hand, this figure demonstrates that drugs act on constructs rather than particular brain regions. On the other hand, it shows that circuits underlying constructs are strongly overlapping but not identical. For any given construct, the underlying circuitry, modulated by interplay between the neural substrates within, flows toward a stable activity pattern – the attractor state **(D)**. The stability means the network will relax to the same stable pattern even after small degree of external stimulation. This phenomenon is pictured with the ball metaphor. Treatment with drugs is most likely to change the patient's state by reshaping the attractor landscape. This drift results in change in the particular construct, whose circuit is targeted with treatment **(E)**, but may also affect other constructs via circuits overlapping with the targeted one. Then, the behavioral consequences of the treatment are the basis to prescribe a more appropriate drug for the given individual **(A)**. VMPFC, ventromedial prefrontal cortex, including BA25 (subgenual cortex); BA9, dorsolateral prefrontal cortex; BA32, dorsal anterior cingulate cortex; BA33, part of anterior cingulate cortex; INS, insula; NAC, nucleus accumbens; CN, caudate nucleus; PT, putamen; THA, thalamus; AMY, amygdala.

dysthymia is significantly different from MDD (16). Third, sometimes new MDD types are distinguished on the basis of specific events triggering the disorder, e.g., grief in the DSM-5 [and in the incoming ICD-11 (14, 17, 18)] and premenstrual dysphoric disorder in DSM-5 (19). This change of diagnostic criteria over time leads to differences in interpretation and is a strong argument for developing an objective approach.

PHYSIOLOGICAL CONSTRUCTS

As mentioned before, there is a variety of competing biomarker sets, each suggesting different MDD etiology. The catecholamine hypothesis of Schildkraut (20), originated in the 60s, advocated that norepinephrine (NE) plays a pivotal role in affective disorders, with a lesser role for epinephrine (E), dopamine (DA), and serotonin (5HT) levels. The hypothesis suggested a reduced level of neurotransmission in E, NE, DA, and 5HT pathways as a possible cause of MDD. Today, it is known that not only DA, NE, and 5HT, but also acetylcholine (AC) has a strong impact on mood (21). Nevertheless, the mechanism of the shift from a healthy brain state into MDD and the role of each of these neuromodulators in this process are not yet understood.

Monoamines and AC are not the only neuromodulatory chemicals involved in MDD. Neuroendocrine mechanisms such as the corticotropin-releasing factor (CRF) may also play a role (22). In depression, this peptide is overproduced in the hypothalamus, which, acting along with arginine vasopressin (AVP), triggers hypersecretion of adrenocorticotrophic hormone (ACTH) from the pituitary. Overproduction of ACTH leads in turn to overproduction of glucocorticoids (cortisol in humans, corticosterone in rodents) from the adrenal cortex. This circuit is known as the

hypothalamic-pituitary-adrenal (HPA) axis, and – as a part of the neuroendocrine system – it controls stress reactions, metabolism, and immunity (23). HPA theory of depression corresponds to the evidence that, due to epigenetic mechanisms, early life events can cause HPA overactivation in adult life (24).

Furthermore, recent observations demonstrate that antidepressant drugs targeting monoamines also modulate synaptic GABA transmission. Additionally, post-mortem studies reveal a dramatic reduction in plasmic GABA concentration in MDD patients. These findings have implicated GABAergic mechanisms in MDD (25), and led to the postulate that the balance of excitation and inhibition (E-I) in brain networks in MDD is disturbed (26).

Another theory of MDD results from the observation that antidepressants induce plasticity in the synaptic strengths, altering patterns of connectivity in the brain (27). Consequently, it was proposed that MDD may reflect a primary impairment in neuronal information processing caused by a disrupted functional or effective (directed) connectivity rather than by any form of chemical imbalance.

NEURONAL CONSTRUCTS

The identification of neuronal circuits underlying MDD with use of fMRI initially has led to the default mode network (DMN) theory of MDD (28, 29). DMN is a circuit defined by slow, coherent oscillatory activity in a wakeful resting state in humans with eyes closed (30). It mostly involves structures engaged in self-referential processes (parts of the medial prefrontal, posterior cingulate and parietal cortices, and medial temporal lobe), as well as the centers for memory (hippocampus, parahippocampal gyrus) and limbic structures (amygdala, nucleus accumbens, hypothalamus) (31). Imaging studies reveal that resting-state activity in many of the DMN nodes is altered in MDD (32). It was recently found that activity in DMN correlates with mood (33), therefore this circuit might be responsible for the affective aspect of the disorder. DMN is just one of many resting-state networks (RSNs) identified so far (34), and methods proposed for identification of MDD on the basis of resting state fMRI respect not only DMN but also other RSNs. For instance, a recently developed computational diagnostic method utilizing Hurst exponent takes into account DMN, right and left fronto-parietal, ventromedial prefrontal, and salience networks (35).

Recent evidence suggests that not only RSNs, but also the central-executive network (CEN) seems to be impaired in MDD (36). This network involves a few subdivisions of prefrontal cortex (PFC), anterior thalamus, and dorsal caudate nucleus. As opposed to RSNs, CEN comes to play during processing that requires cognitive control (37), and therefore is responsible for the executive functions, e.g., response inhibition, reward processing, planning, and working memory. Therefore, as opposed to RSNs, CEN might be involved in such constructs as recurrent thoughts of death and diminished attention. These two families of networks are complimentary and tend to switch the activity between each other.

Identification of common patterns of up- and down-regulation in the nodes of RSNs and CEN could serve as a new, more robust mean to identify network-related biomarkers of MDD (38). In particular, construct-based approach would allow for creating of individual dynamical profiles for patients, and therefore personalized therapy.

TREATMENT

Coming back to causality, we believe that treatments in MDD affect neuronal dynamics, and this dynamics in turn triggers the behavioral change. Treatment choice depends on multiple factors, including the course of the disease, prior medical treatment, etc (39). Evidence-based treatment guidelines suggest cognitive-based therapy [CBT (40)] and pharmacology (41) as the first treatment of choice (42). On the other hand, electroconvulsive therapy [ECT (43)] is only recommended if the aforementioned methods are ineffective for the given patient, whereas deep brain stimulation [DBS (43)], as the most invasive method, is not yet approved by the United States Food and Drug Administration for treatment-resistant depression (43). Even though new treatment methods such as repetitive transcranial magnetic resonance [rTMS, a localized, superficial stimulation of the cortex with magnets (44)] and neurofeedback therapy [a combination of cognitive therapy with neurobiological approach: a real-time feedback of local fMRI signals (45)] are being tested, they are not established methods yet.

An example of drugs as a treatment procedure affecting construct-related circuits, changing the brain dynamical state, and thus influencing the diagnosis is presented in **Figure 1A**.

CIRCUIT FOR MDD

As mentioned in section "Constructs in MDD", our viewpoint is that the mechanism underlying MDD is a superposition of multiple circuits, each of them having a causal effect on one of the cognitive constructs present in MDD. In fact, the number of these cognitive constructs, and therefore also the underlying circuits, may be much higher than the number of diagnostic categories specified in the DSM-5. Exemplary constructs not mentioned in the DSM-5 but present in a vast majority of MDD patients include negative bias in attention and memory (46), a negative view of the world and the future (41), learned helplessness (47), obsessions, and pathological rumination (48).

However, in order to perform a causal inference linking circuits to cognitive constructs, one needs to determine which circuits to study in the first place. MDD is a heterogenous disorder, and, as such, arises from anatomical and functional changes in a wide range of brain regions. The circuits that were first proposed to be responsible for MDD consisted of regions known to be involved in mood. One of these mood generators is the corticomesolimbic loop: one of a few parallel, basal ganglia-thalamo-cortical loops that projects from the ventromedial PFC to the medial dorsal thalamus through the nuclei of the basal ganglia (49). The other mood generator is the aforementioned hypothalamic-pituitary-adrenal axis (HPA) whose dysfunction widely affects monoamine pathways and triggers mood fluctuations. Recently, the viewpoint at MDD and other mental disabilities through the prism of large-scale brain networks identified on the basis of fMRI studies (RSNs and subcircuits of the CEN), and interactions between them, has gained in popularity (50–56).

We take this large-scale perspective. However, as mentioned above, in our view the search for mechanisms underlying MDD

should include zooming into circuits underlying single diagnostic constructs. Large-scale networks are complex and, as such, they might be decomposed into simpler functional circuits. This is definitely the case for the CEN. On one hand, various cognitive constructs could be characterized as different states within the same network. On the other hand, CEN is most probably divided into functional subcircuits which activate while solving particular tasks involving cognitive control, e.g., reward receipt, signal inhibition, decision making, language processing. Another example is the DMN which generates mood. It might be composed of a few interacting subcircuits accounting for generation of basic emotions (57, 58) which do not coexist (59, 60). However, it could also be the case that basic emotions represent various attractors of one large circuit, which is why it is so hard to find specific neuronal underpinnings of basic emotions (61, 62).

In terms of models, so far RSNs are better characterized than CEN (63, 64), probably because of stable temporal dynamics that can be easily investigated with fMRI. Interestingly, Deco et al. (65) propose a model of the resting-state oscillations as a multistable system driven by noise, which is consistent with recent findings on the dynamics of the functional connectivity in RSNs (66–68). It turns out that resting state activity is not uniform but involves numerous modes that switch on and off. Some computational studies suggest that the identified modes of functional connectivity correspond to various eigenmodes of the anatomical connectivity (69), which is a strong argument toward a viewpoint at DMN and other RSNs as a number of interconnected circuits. On the contrary, psychometric studies reveal seven dimensions of cognition during rest: discontinuity of mind, theory of mind, self, planning, sleepiness, comfort, and somatic awareness (70). These dimensions represent various cognitive modes between which subjects switch during the rest. This is an argument on behalf of switching between attractors of one big network during the resting state.

How do the circuits generating single cognitive constructs contribute to this large-scale picture? The construct-wise approach that we take is motivated by circumstantial evidence that, in general, drugs target cognitive constructs rather than the whole disorders. **Figure 1C** presents an example of fluoxetine acting differently in MDD patients with low mood (71–73) and anhedonia (74–76). Influence of fluoxetine treatment on activity in brain areas in positive (77) and negative (78) affect's regime differ (79). On **Figure 1C**, one more phenomenon is demonstrated: circuits underlying constructs diagnostic for MDD are not identical. From comparison of these two simplified circuits for low mood and anhedonia, one can draw a conclusion that some regions are involved in the low mood but not in anhedonia and vice versa. Furthermore, there are regions such as the amygdala that are either up- or down-regulated in MDD, depending on which cognitive construct is present at the moment.

The circuits underlying constructs are overlapping and interacting; however, it seems that – as demonstrated on the example of fluoxetine – pharmacology targets specific constructs rather than the whole disorder. Interestingly, the same drugs are used in mental disorders sharing common cognitive constructs. For example, sertraline is used in the treatment of MDD, obsessive-compulsive disorder, panic disorder, anxiety disorders, post-traumatic stress disorder (PTSD), social phobia, and premenstrual dysphoric disorder, all of them involving fear (80).

MODELING MDD

NEURAL MASS MODELS AND ATTRACTOR LANDSCAPES

So far, psychiatric disorders have not been properly conceptualized in the language of computational neuroscience (81–83). Early research in this field was centered on reinforcement learning models which describe behavior as taking actions which maximize predicted rewards (84). Since DA is believed to be involved in prediction (85, 86), mostly the disorders linked to DA such as schizophrenia were modeled with use of the reinforcement learning (87).

However, since both calculating the odds for possible rewards and taking decisions on the basis of that calculation do not directly correspond to the neuronal activity and physiology of the brain, models based on reinforcement learning are a poor choice when it comes to neuroimaging-based biomarkers for mental disorders. In the last decade, comparing structural and functional connectivity in brain networks in health, in disease, in terms of graph theoretic measures, such as small-worldness (88) or modularity, (89) became a popular research direction (90). These measures have led to multiple interesting results upon the global properties of brain networks in cognitive disorders (91–93) including MDD (94, 95). However, these measures only take undirected connectivity between brain regions into account. The assumption of undirected connectivity yields a conclusion that for every pair of brain regions A and B, once treatment procedure targets region A, it has the same impact on region B, as if one would target region B with the same treatment and measure the change in activity in region A – which is, in general, an unrealistic assumption. Therefore, graph theoretic measures do not extensively incorporate the information that can be rendered from the neuroimaging data and that is of primary importance for assisting diagnosis and treatment in cognitive disorders.

Recently, the concept of attractor networks was proposed, as a tool that might explain cognitive disabilities while corresponding to the neural dynamics in the brain. An attractor network is a network of nodes, often recurrently connected, whose dynamics settle to a pattern stable in time: the so-called attractor state. Analysis of the distribution of attractor states and their basins of attraction, a so-called attractor landscape, was effected on a microscale so far. At the microscale, single neurons are the nodes in the network, and stable firing patterns of those neurons constitute an attractor state (96). This approach is present in contemporary computational neuroscience, e.g., in the models of activity in olfactory (97) and auditory (98) cortices in rodents as well as hippocampal grid cells in humans (99). This concept has also been broadly used in psychiatry. In example, the PFC has been modeled as attractor network in order to explain the deficit in short term memory in schizophrenia (100) and compulsions in obsessive-compulsive disorder (101). Up until now, it is unclear how these models translate to patients because neither the invasive measurements of a single-neuron activity necessary to validate the attractor network models are possible, nor do non-invasive methods have the appropriate resolution.

How about the macroscale? It is now believed that the fMRI research can provide the insight necessary to understand cognitive constructs (102, 103). But is the concept of attractors also applicable for this sort of data? Here, we propose a conceptual advance to apply mathematical modeling directly to patients. This proposal involves looking at the large-scale neural circuits in order to perform attractor landscape analysis on the macroscale. Mind that brain circuits are networks of interacting nodes, and therefore can be represented and analyzed as dynamical systems, in a similar fashion as networks of single neurons. As opposed to microscale, at the macroscale whole brain areas account for the nodes in the network, and attractor states are stable activity patterns across all nodes within the network. For example, in case of the fMRI data, the overall activity in a region of interest can be expressed as the summation over activity of all voxels within that region. This data is very convenient for neural mass models when it comes to modeling cognitive architectures (104). The principal idea of neural mass models is setting the density of neurons to the continuum limit in modeling the activity of large neural populations. This assumption of spatially continuous neural networks thus allows for analytical treatment of such global variables as firing rate in space and time. An example is the classic Wilson–Cowan mean-field model (105). In this model, the activity of neuronal populations (or brain regions) is represented by dynamical variables. **Figure 2** presents a simplified version of the model where spatial patterns of spiking activity are replaced by one dynamical variable. In the model, effectively connected neuronal populations, representing brain regions, interact and are additionally tuned by neuromodulators. Such dynamical systems have a number of stable attractors, and therefore a number of basins of attraction. The possibility is that in MDD patients, the shape of the attractor landscape for a particular cognitive construct is different than in healthy controls. However, it can also be that they occupy a "wrong" attractor state (106).

TREATMENTS IN THE CONTEXT OF DYNAMICAL SYSTEMS

All of the available treatments affect the dynamics of large-scale networks and therefore also the attractor landscapes (108–110). Therefore, with use of the Wilson–Cowan model, one can then investigate the landscape of basins of attraction in response to the treatment procedures. Antidepressant drugs can reshape the attractor landscape in multiple ways: they can lower the hills of the landscape around the current state of the patient or make the current attractor state shallower in order to facilitate escaping from the local minimum (**Figure 2C**, upper). The drugs can potentially also modify background neuronal noise, which in turn may affect the probability of occupying different attractor states (111). On the other hand, stimulation methods that regulate the neural dynamics directly, such as rTMS, ECT, and DBS can influence the state of the patient by providing a brief pulse to the brain network in the patient and thus allowing the brain network to leave the "wrong" attractor state immediately (**Figure 2C**, lower). Interestingly, in the treatment-resistant depression, electrical stimulation through ECT and DBS prove to be highly effective (112, 113), which means that, under some circumstances, they perform better than drugs, or even than the cognitive therapy which targets the cognitive constructs directly. This provides some hint suggesting that looking at clinical symptoms of MDD through the prism of neuronal circuits, and targeting treatments at those circuits might be more beneficial than any other treatment, including, paradoxically, even the behavioral treatment centered at specific cognitive traits in MDD.

CIRCUIT TO CONSTRUCT MAPPING

WHAT IS CCM

Every patient has a different, individual attractor landscape. This landscape reflects such personal traits as the size of the brain regions involved in MDD, functional connectivity within DMN and CEN, baseline concentrations of monoamines, and all the other endogenous chemicals that influence the excitation-inhibition balance in the brain. During rest, DMN and other RSNs are active and the patient occupies stable attractors in their attractor landscapes. On the contrary, during solving cognitive tasks, subnetworks of CEN come to play (depending on the nature of the task) and the brain state jumps to one of its (most probably, also stable) attractors. We predict that a disturbance of the attractor landscapes within the DMN should account for the cognitive constructs involving affective components of MDD, whereas disturbance of the attractor landscapes within cognition-related RSNs (such as fronto-parietal network) and within the CEN should be responsible for the cognitive constructs involving executive functions.

But how do these attractors map onto cognition? Let us consider a brain network consisting of interconnected nodes described by their activities, either in resting state or in some cognitive process (**Figure 3**). While looking for causal interactions between neuronal circuitry and behavioral outcome, one should perform a mapping from a multidimensional space spanned by patterns of neuronal activity (namely, attractors of the neuronal networks) onto a multidimensional space spanned by the cognitive constructs. This is what we called the CCM approach. The direction of causal inference in CCM goes from circuitries toward behavior because the CCM approach is designed for better treatment, which should ultimately target the diagnostic cognitive constructs in MDD. Therefore, it is essential for the constructs to be compact, but the underlying circuits can be complex as is necessary.

The CCM approach involves performing this mapping with use of joint imaging and psychometric methods on large clinical datasets. Once we identify the circuits underlying single cognitive dimensions of MDD, we can perturb this construct-related circuits in a single patient with treatments, affecting the neuronal dynamics, and tracking both the resulting position in cognitive construct space and the dynamical properties in the construct-related circuits.

EXECUTION OF CCM

Execution of CCM is a multistep process. The preliminary step is to determine an extensive list of constructs involved in MDD. Since the classic diagnostic tools are questionnaires and experimental tasks, this analysis would run through a number of various variables, grouping them into dimensions, with a subsequent sanity check if the outcome constructs have a consistent content. The list of constructs determined in this protocol can be longer than the list of the DSM-5 criteria, thus we call the constructs with

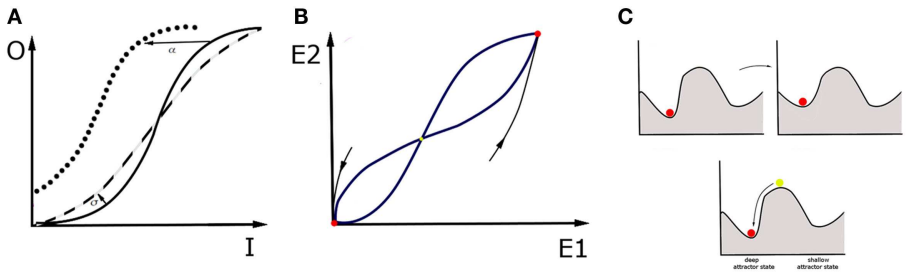


FIGURE 2 | Wilson–Cowan model and a "ball" metaphor. The activity of a single brain area within the network is a consequence of the synaptic inputs from other areas, the modulatory tone generated by diffuse projections, and the recurrent connectivity within the brain area itself. The activity reflects a specific balance between excitation and inhibition within the area. For simplicity, we describe the activity by one variable, *E*, for which the following equation holds: $\tau\frac{dE}{dt} = -E + f(\alpha E + \beta I + \gamma M)$. The first term on the right tells us that in the absence of any drive (provided by the second term), the activity decays to zero with time scale τ. The second term incorporates the contribution of recurrent connectivity via *E* itself, input from other areas, represented by *I*, and the level of neuromodulation, represented by *M*. Each of these contributions are weighted by factors: α, β, and γ respectively. When the second term is positive, it increases the level of activity. The function *f* is a response function that translates the sum of activities into a driving term, and is typically sigmoidal (106): $f(x) = \frac{Ax^2}{x^2+\sigma^2}$. In this form, *A* is the maximum that *f* can reach for large *x* values, and σ is the value for which *f* is equal to half its maximum value. In addition, it also specifies how steeply *f* increases with *x*, a quantity that is also referred to as the gain factor. Note that this expression only holds for positive *x* values, *f* is zero when *x* is negative. This model has a range of parameters, which is important because each of them can be linked to specific physiological processes and changes in circuit structure. For instance, an increased β represents a stronger synaptic projection, whereas an increased α represents stronger recurrent synapses. An increased *M* reflects the effect of neuromodulators that increase the level of depolarization in the cells, and hence the baseline firing rate; γ reflects the sensitivity to neuromodulators of cells and circuits. The value of σ can be interpreted as a change in gain. **(A)** In a given region, the sigmoidal input-output (I-O) relationship has three regimes. For small input $y << \sigma$, it increases rapidly. For large inputs, $y >> \sigma$, it saturates. For values in between, it connects these regimes linearly. If the σ value, and thus excitability of the region, grows (dashed line), the I-O function is steeper than in the control case (solid line). If the region gets stronger recurrent connectivity, input from other regions or neuromodulation, so that the α, β, γ values grow respectively, I-O function shifts to the left (dotted line). **(B)** In an example of two interconnected regions, E_1 and E_2, this dynamical system has three fixed points that are candidates for attractor states. In this example, two of them are stable (red). For a given attractor, setting activities E_1, E_2 to arbitrary initial values within the basin of attraction will make the system move on toward this attractor. The third fixed point is unstable (yellow), which means that every small perturbation from this state makes the system fall into one of the basins of attraction, and thus end up in one of two attractor states. **(C)** One may picture attractor states with the ball metaphor. Disease can be represented in two ways. It can mean a change in the landscape of basins of attraction: some attractor states change position and even if the patient occupies the original attractor throughout the process, their brain state gradually changes the attractor state that they occupy. This can be achieved by changing shape of I-O function with use of parameters σ and α, β, γ or changing of relaxation time constants τ. However, it can also mean that, in a result of intrinsic noise in the brain or in response to a particular external input, the brain state in the patient is triggered to switch to another "wrong" basin of attraction. The noisy behavior of the network is not captured by the basic version of Wilson–Cowan equations, but incorporating noise in and therefore also a stochastic driving force is also possible. An attractor is a network state where the levels of activity do not change anymore, hence *E* is constant. Mathematically, this means that *E* does not change over time, hence that its value is given by setting the right hand side of equation (1) to zero, which yields $E = f(\alpha E + \beta I + \gamma M)$, hence *f* gives the steady state values, hence increases in the factors α, β, and γ immediately increase the *E* value. It is important to realize that this is an equation from which *E* needs to be found. In the preceding, we focused on a single variable *E*, but in a network there is at least one variable for each brain area involved. For multiple brain regions involved, which is true in MDD, $\tau_i\frac{dE_i}{dt} = -E_i + f_i\left(\sum_j J_{ij}E_j + \gamma_i M_i + I_{\text{stim},i}\right)$. Here *i* represents the index of the brain area and *j* is the index of brain areas that provide input. Most parameters now have an index *i*, because their value depends on the area they represent. We have also included a stimulation current, which represents the effects of electric or magnetic stimulation. Within this framework, the effects of treatments can be captured. On one hand, treatments can reshape the attractor landscape. For instance, pharmacological manipulations can either change the level of neuromodulation or the sensitivity of the circuit to neuromodulators. This would lead to the homeostatic regulation of the coupling coefficients J_{ij}, and σ, and, subsequently, to the change in the map of attractors. On the other hand, a single electrical stimulation, such as ECT session, could change the attractor, offering temporary relief; but if the new attractor is not stable, the brain network could return to the old attractor over time. A sequence of electrical stimulation would also affect J_{ij} and thus change which attractors are possible and how stable they are. Taken together, electrical stimulation has the advantage that its effect is local and can be tuned to alter/correct a specific J_{ij} value.

anonymous t_i in the **Figure 3A**. Furthermore, some constructs may be heritable and thus fulfill the definition of endophenotypes, which is especially relevant for executive functions (114), whereas other constructs such as recurrent thoughts of death are not likely to be heritable. However, this analysis will not reveal whether a given construct is heritable or not.

The second step is to find neuronal mechanisms of each of the obtained constructs. For every single construct, one should start the procedure from the first order analysis: investigating patterns of activation and effective connectivity in a cohort of patients exhibiting that construct (and, of course, a cohort of controls), in order to identify the underlying neuronal network and to build a corresponding dynamical system (**Figure 3A**). Using Pearl's definition of causality, for the effect size large enough we can determine causal effects on the basis of this observational study.

If this first level analysis does not identify unique circuitry, there can be multiple interacting circuitries involved in the construct. In that case, one should perform a second order analysis. For instance, one can perform repeated diagnostic evaluation and repeated fMRI imaging assessment longitudinally within the same patient. Then, using autoregressive models in order to analyze the time course of the construct and correlating these independent components with neuroimaging data should reveal independent components in the circuitry underlying this construct.

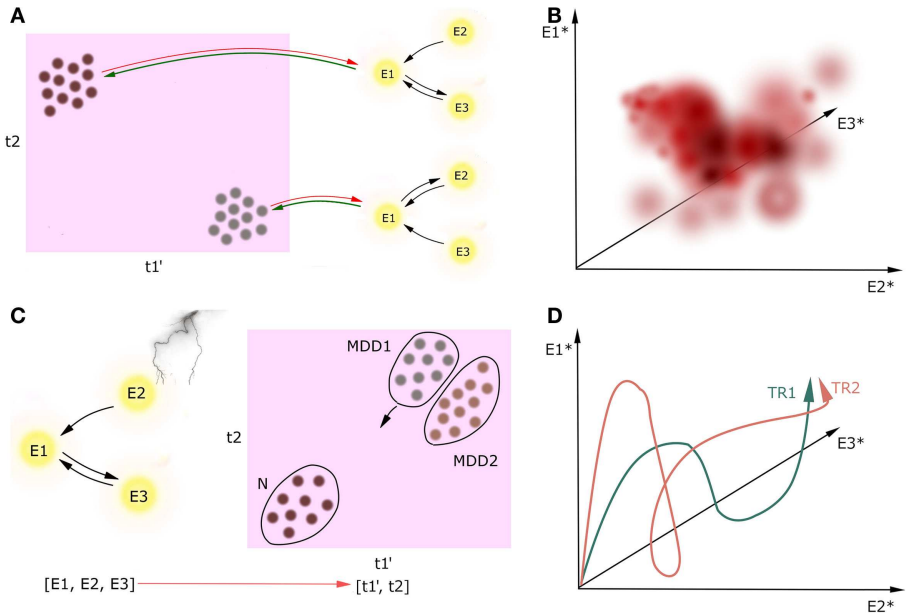


FIGURE 3 | Circuit to construct mapping. Causality between activity in underlying network (nodes E1, E2, E3) and the multidimensional construct space (t1, t2) can go two ways, but we are only interested in neural circuitry as the cause and cognitive constructs as the effect [**(A)**, green lines]. Behavioral learning and neuroplasticity can give the backward direction of causality [**(A)**, red lines]; however, we do not cover this issue in this paper. We refer to cognitive constructs as t_i because the preliminary step of the CCM includes determining the full list of the involved constructs which can be broader than the list of the DSM-5 diagnostic criteria. Prime denotes endophenotype, dot denotes a patient, asterisk denotes attractor state. Firstly, circuitries involved in these constructs should be linked on the basis of extensive observational research on a large cohort of patients **(A)**. Secondly, we can plot how these attractors map onto cognitive constructs with temperature maps **(B)**. In this example, we plot the value of one continuous construct (which can represent, i.e., level of mood) in a three-dimensional space spanned by the attractors of the underlying three-node network. Thirdly, one can track the current state of one patient both in the multidimensional construct space during the treatment [**(C)**, a scatter plot in case we want to track multiple patients at a time]. Distribution of patients in this space may reveal subtypes of MDD (MDD1, MDD2). Moreover, when the network manipulation with treatment is sufficient, it can trigger the patient's brain state toward a new attractor in some of the construct-related dimensions, and in a result the patient flows to another point in the construct space. Since we will create these maps on the data from the limited number of patients in the cohort, this temperature maps will not span the whole volume of possibilities. Lastly, one may investigate how the treatments TR_i affect the attractors of the underlying networks **(D)**. In this case, we have three nodes in the underlying networks, which means that the attractors of this system will be points in a three-dimensional attractor space.

We predict that positive correlations between revealed cognitive constructs across patients are inevitable, which should be reflected in overlaps between circuits underlying the constructs. We can also analyze how the attractors of the dynamical systems map onto cognitive constructs using temperature maps (**Figure 3B**). Since we will create these maps on the data from the limited number of patients in the cohort, this temperature maps will not span the whole volume of possibilities.

The third step is building the dynamical models representing the identified circuitries underlying cognitive constructs. The proposed Wilson–Cowan model can be applied to any clinical data that reveals the distribution of activity in the brain over time (115), in particular to blood oxygen level dependent (BOLD) signal in fMRI (116) or EMG/EEG data (117). Wilson–Cowan model has some similarities to the dynamical causal modeling (DCM), a well established method for extracting effective connectivity for both fMRI and EEG/EMG data (118–124), in a sense that it describes the neuronal communication between brain regions in terms of ordinary differential equations. The major difference is that – in both classical (119) and recent stochastic version of DCM for fMRI data (125) – there is an assumption of linear transfer functions, whereas it is known that large neuronal populations exhibit sigmoidal rather than linear response to the external inputs (106), which is incorporated in the Wilson–Cowan equations (126).

In this procedure, a single patient in a cohort is just an object to the explanatory science. However, once the circuitries underlying cognitive constructs involved in MDD are determined, the patient may become a subject in a case study, and receive a personalized treatment. Investigation of the trajectory of the particular patient in the construct space in response to changes in the circuit activity caused by treatments (**Figure 3C**) might not only provide new biomarkers for MDD and better insight into the mechanisms of treatments, but also answer the question of how to predict resilience to treatment. This research may also elucidate factors that determine whether a treatment is effective to a particular group of patients. Furthermore, this analysis might help to address the question if the mental disorders of interest, e.g., MDD, are homogenous or split into subtypes on the basis of the patient trajectories in the construct space. Lastly, one may investigate how

the treatments TR_i in the given patient affect the attractors of the underlying networks (**Figure 3D**).

BENEFITS OF CCM

Circuit to construct mapping brings three new qualities to the table. Firstly, treating networks as dynamical systems allows one to extract and to characterize global properties of the networks involved in cognitive constructs in a comprehensive and versatile way. So far, research in human imaging was focused on finding particular areas involved in cognitive tasks by virtue of stable activation patterns, or investigating context-dependent strength of connectivity between particular areas. These are two out of many viewpoints which one can take in order to characterize large-scale brain networks. In fact, these are the two sides of the same coin: the distribution of activation patterns in a network is a global property emerging from behavior of the underlying dynamical system specified through the connection strengths between areas. Whether the activity patterns are more informative than the connectivity strengths, depends on the circumstances. In **Figure 4A**, we present a toy example. Let us assume that, in the simplest case, our sigmoidal transfer function can be approximated as a linear function. For some combinations of inputs to the network and connection weights, a small change in connection weights (by 10%) yields an enormous change in the value of stable activation patterns (by 1000%, upper panel). For other combinations of weights and inputs, even huge change in the connectivity strengths (by 300%) yields a small change in stable activation patterns (by 10%). As a consequence, whether activity patterns in the networks are sensitive to changes in connectivity strengths depends on the tuning in the network, for instance on the balance between connectivity weights in the network and external conditions such as experimental inputs. Therefore, since the dynamical systems incorporate both connectivity (as the cause) and about stable activity patterns (as the effect), they integrate the two sorts of information about the circuits into one framework.

Secondly, the decomposition of psychiatric disorders into a number of diagnostic traits allows for fundamental explanatory research in psychiatry, and therefore also for new, neuroimaging-based biomarkers for cognitive disorders. In terms of causal modeling, gathering clusters of traits into big cognitive paradigms such

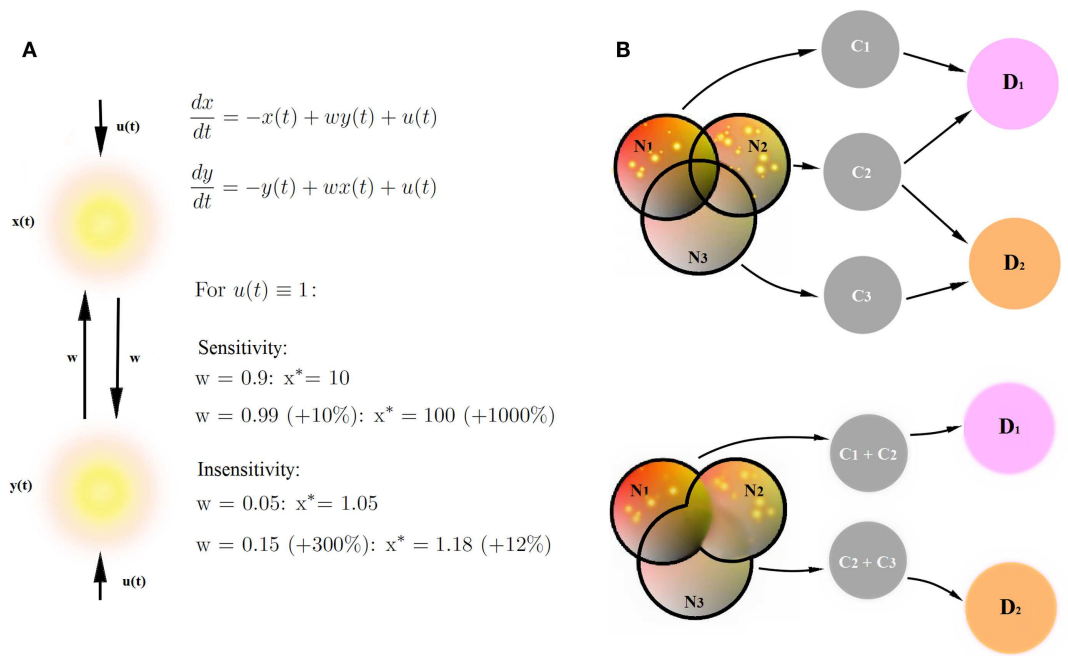


FIGURE 4 | Benefits of CCM. **(A)** Stable connectivity patterns and connectivity strengths are two sides of the same coin. Let us assume that our sigmoidal transfer function can be approximated as a simple linear function. We provide an example of a network sensitive to changes in connectivity strengths (a small change in connection weights yields an huge change in the value of stable activation patterns) and an example of a network insensitive to changes in connectivity strengths (a huge change in the connectivity strengths yields a small change in stable activation patterns). Therefore, the description of networks by means of dynamical systems provides more versatile description than connectivity strengths in the networks or stable connectivity patterns alone. **(B)** Decomposition of psychiatric disorders into a number of diagnostic traits helps causal inference in the diagnostic process. Networks N_i underlie cognitive constructs C_i diagnostic to psychiatric disorders D_i. Disorders D_1 and D_2 share a common cognitive construct C_2, but involve also disorder-specific diagnostic constructs C_1 and C_3. Let us assume that the joint posterior distribution for every disorder D_i factorizes into posterior probability distributions for single diagnostic constructs. Let us further assume that we find the same, specific pathologies in networks N_1 and N_2. If we can decompose the D_1-related network into a sum of networks N_1 and N_2 underlying single diagnostic constructs C_1 and C_2 (upper panel), we collect more evidence for the disorder D_1 (pathologies in both networks linked to diagnostic constructs) than for the disorder D_2 (pathology in one out of two networks linked to diagnostic constructs). On the other hand, if we are not able to decompose D_1 and D_2-related networks into networks underlying single diagnostic categories (lower panel), the amount for evidence in favor of both disorders is the same because both networks $(N_1 + N_2)$ underlying disorder D_1 and $(N_2 + N_3)$ underlying disorder D_2 are pathological.

as psychiatric disorders can be misleading, given that the disorders strongly overlap in terms of diagnostic criteria. A simple example is provided in **Figure 4B**. In this example, overlapping networks N_i underlie cognitive constructs C_i, which are diagnostic to psychiatric disorders D_i. Disorders D_1 and D_2 share a common cognitive construct C_2, but involve also disorder-specific diagnostic constructs C_1 and C_3. In this toy example, let us assume that the prior probabilities of cognitive constructs C_i are equal and that likelihood of the pathologies in networks N_i given constructs C_i are the same. Let us further assume that in our patient, we find the same, specific pathologies in networks N_1 and N_2. If we can decompose the D_1- related network into a sum of networks N_1 and N_2 underlying single diagnostic constructs C_1 and C_2 (**Figure 4B**, upper panel), we can perform statistical inference, linking specific changes in N_1 and N_2 with constructs C_1 and C_2, respectively, and collecting evidence behind the hypothesis that the patient is a subject to the disorder D_1. Since C_2 is also a construct diagnostic to the disorder D_2, we also collect some evidence behind the hypothesis that the patient suffers from the disorder D_2. However, assuming that the joint posterior distribution for every disorder D_i factorizes into posterior probability distributions for single diagnostic constructs, we collect more evidence for the disorder D_1 than for the disorder D_2.

On the other hand, if we are not able to decompose D_1 and D_2-related networks into networks underlying single diagnostic categories (**Figure 4B**, lower panel), the amount for evidence in favor of both disorders is the same because both networks $(N_1 + N_2)$ underlying disorder D_1 and $(N_2 + N_3)$ underlying disorder D_2 are pathological, and we are not able to extract any disorder-specific subnetworks which would provide any further evidence in favor of one of the disorders. Therefore, decomposing mental disorders into single diagnostic constructs and linking construct-specific circuits is of primary importance for explanatory models in psychiatry.

Thirdly, CCM as a modeling procedure that projects neuronal dynamics straight into behavioral dimensions of MDD, could not only serve as explanatory model when applied to a large cohort of patients, but also enhance the current treatment selection for individual patients and make a step toward the personalized medicine. In order to perform explanatory research "in Pearl's sense," we need to use neuroimaging along with behavioral data from a large cohort of patients because, in order to reveal the circuitries underlying MDD-related cognitive constructs, we need to find systematic differences in circuit dynamics that result in systematic differences in behavior. But once this explanatory research is done and the circuitries underlying cognitive dimensions of MDD are defined, zooming into the circuit dynamics and its development under treatment in a particular patient would allow for the personalized interventions.

LIMITATIONS OF CCM APPROACH

PLASTICITY AND NEURODEGENERATION

So far, sensory systems are best characterized in terms of underlying circuitries. However, events in sensory systems happen on a millisecond to second timescale whereas the evolution of psychiatric disorders is a few orders of magnitude slower and therefore might be much more complex. MDD may result from traumatic experience or emerge without a particular inducing event, but in any case the process of falling into a depressive episode lasts for weeks, as opposed to perceptual learning which takes only seconds. Also, some treatment procedures are long lasting, i.e., MDD pharmacotherapy is primarily monoamine based and typically requires intake for 3–4 weeks prior to symptomatic improvement (with the exception of ketamine). This time course is a major impediment to modeling MDD because imbalance in mood may arise not only on top of changes in neurotransmitter concentrations, but also result from other processes such as structural plasticity and neurodegeneration (127). The mechanisms underlying these two processes are not fully understood, and, in the case of structural plasticity, is difficult to investigate in a living human brain. Neural mass models can only serve to compare between different stages of the disorder in an individual, and between different individuals at the same stage, yet does not provide a framework that demonstrates real-time evolution of MDD.

HETEROGENEITY

MDD is a heterogenous disorder. The diagnostic criteria are still evolving, and the recently published DSM-5 diagnostic criteria for MDD allow for a variety of diagnostic combinations of cognitive constructs. Is there a plethora of different MDD types, or rather one prevalent state of mind that manifests itself in various ways depending on the patient? This remains an open question. Furthermore, in the literature, there is often no clear distinction between patients who experience a first depressive episode and those who suffer from recurrent depression whereas, as neurodegeneration proceeds and the severity of symptoms elevates, the course of the disease plays the crucial role in the treatment procedure. This also provides a hindrance to the modeling procedures since the information about the stage of the disease is often missing from databases.

Furthermore, complexity of MDD might project also to strongly overlapping construct-related circuits. In example, it was found that the same brain area may host different circuits, which, when activated, have opposing effects on anxiety (128). Furthermore, fMRI studies reveal anticorrelated networks to be activated during cognitive tasks (129). This is circumstantial evidence that multiple distinct circuits can underlie single cognitive constructs (**Figure 1C**). Furthermore, the same constructs can arise from different mechanisms. In **Figure 5**, we discuss impairment in maintaining attention as an exemplary construct that may develop in the PFC of the MDD patients from distinct processes.

APPLICATION OF TREATMENTS TO THE CCM

Some of the possible applications of CCM such as DBS and ECT require invasive methods that cannot be used in humans on a daily basis, and thus require rodent models. Rodent models of MDD are a well explored discipline. However, whether rodent models in mental disorders are fully translational remains unclear, which presents another difficulty for modeling studies. Whereas anhedonia, weight loss and gain, hypersomnia, or psychomotor retardation can be measured in a rodent, some other constructs such as the presence of recurrent thoughts of death, have no equivalent in rodents. On the other hand, modeling that requires invasive techniques such as electrophysiology cannot be ethically introduced

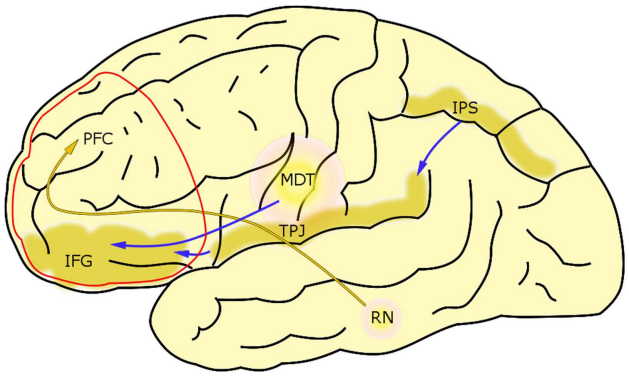


FIGURE 5 | Attention as an example of a construct with multiple neural mechanisms underneath. Maintaining attention can be disrupted by at least two distinct mechanisms: (1) Oversensitivity of the ventral attention network. Imaging studies revealed two systems managing attention in humans. On one hand, we have dorsal attention system, consisting of frontal eye fields (FEF) and intraparietal sulcus (IPS), controlling voluntary deployment of attention (top-down control). On the other hand, we have a right-lateralized ventral attention network (VAN), responsible for orienting attention toward sensory stimuli. It involves temporoparietal junction (TPJ), intraparietal sulcus (IPS) in the parietal cortex, and inferior frontal gyrus (IFG). IFG, as a part of orbitofrontal cortex, receives a strong excitatory input from medial dorsal thalamus (MDT). Since MDT is overactive in MDD, this effect can make ventral attention network oversensitive to stimuli, and as a result holding attention on salient stimuli becomes difficult to the patient. (2) Diminished communication through coherence in the prefrontal cortex. Serotonin produced in the raphe nucleus (RN) modulates gamma oscillations in the prefrontal cortex (PFC), most probably by acting on fast-spiking interneurons expressing serotonin 5-HT_2 and 5-HT_6 receptors. Gamma oscillations play a key role in higher cognitive processes, including attention and working memory. Since serotonergic input to the prefrontal cortex is known to be diminished in MDD, the decrease in gamma power may account for the effect of distractibility in MDD. Both of the above mechanisms lead to a decrease in inhibition within the prefrontal cortex, which might explain why the attention, managed in the PFC, both can be disrupted in a result of hyperactivity of the medial dorsal thalamus and hypoactivity of the raphe nucleus.

into living human brains except under certain prescribed neurosurgical situations. However, the TMS-, pharmacotherapy- and neurofeedback-related CCM approach constitutes an adjunct to rodent models and, as a non-invasive method, it is applicable to patients. Among the emerging treatment methods, neurofeedback seems to be a promising therapeutic procedure for CCM. This method is known to change connectivity in the functional networks (130, 131), but its mechanisms of action are not yet known. Yet the concept of guided self-modulation in a patient in absence of any third-party tools such as electric current or drugs is tempting. However, CCM can also be paired with all the other treatment procedures.

What can be a hindrance in application of the pharmacotherapy-related CCM is that it is difficult to target a given construct with a particular drug because MDD drugs act on monoamine receptors, which are ubiquitous in the brain and present in multiple circuits at a time (**Figure 1C**). Furthermore, some brain regions are hubs that are affected in many constructs thus, targeting these nodes with any form of treatment will have broad consequences for the global brain state. For example, the ventral medial PFC is a major hub in the limbic system known to be involved in low mood (72), anhedonia (75), feelings of worthlessness (132), and diminished working memory (133) in MDD. However, the idea is to provide the online readout for the dynamics of all the involved circuits at a time. Due to this approach, the clinician may first apply a specific treatment in order to target a desired cognitive construct, and then observe how the other construct-related circuits evolve along with the targeted one.

TEMPORAL DYNAMICS IN THE RESTING STATE

Circumstantial evidence suggests that in some aspects, MDD might require deeper insight into activity of neural networks than the afforded by global patterns of activity in the populations of brain regions as obtained from fMRI studies. For example, the DBS has different remission rates depending on the temporal characteristic of the applied current. As it was recently demonstrated that in the Parkinson's disease, temporally irregular DBS is more effective than oscillatory stimulation (134). This effect suggests that in addition to the modulatory effect on E-I balance, electrical stimulation can change the communication between the targeted region and its efferents by affecting communication through coherence (135). This means that the fMRI data, as they are lacking the temporal characteristics in the brain activity, might give an incomplete information about mechanisms of MDD. However, CCM is still a substantial progress for the therapy and treatment in mental disorders, and gives a first insight into the circuits involved in the disorder that opens possibilities for further, more in depth research.

EFFECTIVE CONNECTIVITY IN EEG/EMG AND fMRI RESEARCH

So far, there are papers whose authors use Ising models in order to provide a global description of network properties (as a number of so-called patterns stored in the network (136). However, Ising models are defined only for undirected networks and, in order to use full potential of the CCM, this approach needs a step further by making connectivity directional. In fMRI research, parcelation of the brain into regions is quite successful (137; Oort, in preparation); however, determining connectivity strengths between the nodes is harder because of the limited amount of the temporal information in the fMRI data. So far, the only widely used inference procedure for effective connectivity on the basis of fMRI data is the aforementioned DCM; however, it is only applicable for very small networks 3–4 nodes, requires predefinition of a number of parameters and of network nodes, and in addition to that, as an inference procedure, encounters some critics in the field (138). Since region definition in causality for fMRI is extremely important (Bielczyk et al., in preparation), there is an urge for new, more data driven methods for approaching effective connectivity in these datasets.

In the field of EEG/EMG on the contrary, the problem of causality is orthogonal to the fMRI field: the DCM procedure is quite successful in finding effective connectivity between the nodes of the network, however the optimal method for defining the nodes as sources of the potentials recorded on the scalp is still an open problem. Three popular approaches are dipole modeling, dynamic imaging of coherent sources and frequency-domain minimum

current estimation (139). These methods successfully identify the main sources of oscillations in the brain volume, however there is a room for improvement in terms of the spatial resolution of reconstructed sources.

CONCLUDING REMARKS

As proposed by RDoC initiative, symptoms diagnostic for psychiatric disorders should be interpreted as psychopathological constructs, which need to be investigated, diagnosed, and treated independently. The CCM approach addresses this demand, and provides with a new outlook at clinical treatments in mental disorders. Namely, the treatments not only regulate levels of neuromodulatory substances but also change the dynamical state of the brain by regulating excitation-inhibition balance across brain circuits, which can be tracked with neuroimaging. This change in dynamics may be achieved in two ways: by inducing the structural and functional plasticity that changes the functional connectivity in the circuit (through drugs), or by providing stimulation/inhibition to discrete circuit node (s) and therefore changing the global balance in the brain (through electrical stimulation).

In this work, we underscore the potential of computational modeling in psychiatry as a tool to unravel mechanisms underlying the diagnostic symptoms, to cluster diagnostic cohorts and to customize approach to clinical populations in psychiatry. In addition to this, we anticipate that in the near future, new, personalized treatment methods based on non-invasive regulation of specific neuronal populations' activity with gene therapy may be possible. This approach is still in its infancy and remains to be clinically validated. However, gene therapy up-regulation of p11 protein in the rodent nucleus accumbens proved to cause a reversal of an anhedonic phenotype (140).

Due to our assumptions, diagnostic symptoms of MDD are caused by (mal)behavior of the underlying neuronal circuits. Therefore, we suggest that clinical groups homogenous in the circuit dynamics should also be responsive to similar treatments. Conducting the diagnosis in terms of circuit defects based on the construct domain will then ensure the clinical groups are clustered, and represent more homogenous groups. Furthermore, comparison of depressed patients and healthy controls in the construct space may assist in the investigation if MDD is a single disorder (and diagnostic category) or whether it should be split into diagnostic subtypes. It may also reveal cognitive and neuronal signatures of the phenomena of treatment-resistance. Tracking patient's position in the construct space in response to stimulation/inhibition on one hand, and the evolution of relevant attractor landscapes on the other hand, may provide new insight into the nature of treatments and help to create personalized medicine.

AUTHOR CONTRIBUTIONS

Collecting materials: NB. Drafting of the manuscript: NB. Critical revision of the manuscript, clinical part: JG, JB. Critical revision of the manuscript, computational part: PT.

ACKNOWLEDGMENTS

We would like to thank Dr. Raoul-Martin Memmesheimer, Department for Neuroinformatics, Radboud University Nijmegen, Nijmegen, Netherlands for advice concerning the computational paradigm and Dr. Maarten Mennes, Department for Cognitive Neuroscience, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands for consulting the application of imaging methods. We would also like to thank the Reviewers for constructive and insightful comments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°278948.

REFERENCES

1. Nestler EJ, Barrot M, DiLeone RJ, Eisch AJ, Gold SJ, Monteggia LM. Neurobiology of depression. *Neuron* (2002) **34**(1):13–25. doi:10.1016/S0896-6273(02)00653-0
2. WHO, editor. *Depression: A Global Crisis*. 20th Anniversary of World Mental Health Day. Vienna (2012).
3. Olesen J, Gustavsson A, Svensson M, Wittchen HU, Jönsson B, CDBE2010 study group, et al. The economic cost of brain disorders in Europe. *Eur J Neurol* (2012) **19**(1):155–62. doi:10.1111/j.1468-1331.2011.03590.x
4. Hong CJ, Tsai SJ. The genomic approaches to major depression. *Curr Pharmacogenomics* (2003) **1**(1):67–74. doi:10.2174/1570160033378295
5. Vialou V, Feng J, Robison AJ, Nestler EJ. Epigenetic mechanisms of depression and antidepressant action. *Annu Rev Pharmacol Toxicol* (2013) **53**(1):59–87. doi:10.1146/annurev-pharmtox-010611-134540
6. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. The epidemiology of major depressive disorder: results from the national comorbidity survey replication (ncs-r). *JAMA* (2003) **289**(23):3095–105. doi:10.1001/jama.289.23.3095
7. Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* (2003) **160**(4):636–45. doi:10.1176/appi.ajp.160.4.636
8. Dubin M, Weissman M, Xu D, Bansal R, Hao X, Liu J, et al. Identification of a circuit-based endophenotype for familial depression. *Psychiatry Res* (2012) **201**(3):175–81. doi:10.1016/j.pscychresns.2011.11.007
9. First MB. *The National Institute of Mental Health Research Domain Criteria (RDoC) Project: Moving Towards a Neuroscience-Based Diagnostic Classification in Psychiatry*. New York, NY: Oxford University Press (2013).
10. Lewis D. Causation. *J Philos* (1973) **70**:556–67. doi:10.2307/2025310
11. Holland PW. Statistics and causal inference. *J Am Stat Assoc* (1986) **81**(396):945–60. doi:10.2307/2289069
12. Woodward J. *Making Things Happen*. Oxford: Oxford University Press (2003).
13. Bollen KA, Pearl J. Eight myths about causality and structural equation models. In: Morgan SL, editor. *Handbook of Causal Analysis for Social Research*. Dordrecht: Springer (2012). p. 301–28.
14. Kupfer DJ, Regier DA. Neuroscience, clinical evidence, and the future of psychiatric classification in DSM-5. *Am J Psychiatry* (2011) **168**(7):672–4. doi:10.1176/appi.ajp.2011.11020219
15. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization (1992).
16. Cristancho M, Kocsis J, Thase M. Dysthymic disorder and other chronic depressions. *Focus* (2012) **10**(4):422–7. doi:10.1176/appi.focus.10.4.422
17. Maj M. Bereavement-related depression in the DSM-5 and ICD-11. *World Psychiatry* (2012) **11**(1):1–2. doi:10.1016/j.wpsyc.2012.01.001
18. Mateen FJ, Dua T, Shen GC, Reed GM, Shakir R, Saxena S. Neurological disorders in the 11th revision of the international classification of diseases: now open to public feedback. *Lancet Neurol* (2012) **11**(6):484–5. doi:10.1016/S1474-4422(12)70125-4
19. Epperson CN, Steiner M, Hartlage SA, Eriksson E, Schmidt PJ, Jones I, et al. Premenstrual dysphoric disorder: evidence for a new category for DSM-5. *Am J Psychiatry* (2012) **169**(5):465–75. doi:10.1176/appi.ajp.2012.11081302
20. Schildkraut JJ. The catecholamine hypothesis of affective disorders: a review of supporting evidence. *Am J Psychiatry* (1965) **122**(5):509–22. doi:10.1176/ajp.122.5.509
21. Warner-Schmidt JL, Schmidt EF, Marshall JJ, Rubin AJ, Arango-Lievano M, Kaplitt MG, et al. Cholinergic interneurons in the nucleus accumbens regulate depression-like behavior. *Proc Natl Acad Sci U S A* (2012) **109**(28):11360–5. doi:10.1073/pnas.1209293109

22. Pariante CM, Lightman SL. The HPA axis in major depression: classical theories and new developments. *Trends Neurosci* (2008) **31**(9):464–8. doi:10.1016/j.tins.2008.06.006
23. Sánchez MM, Ladd CO, Plotsky PM. Early adverse experience as a developmental risk factor for later psychopathology: evidence from rodent and primate models. *Dev Psychopathol* (2001) **13**(3):419–49. doi:10.1017/S0954579401003029
24. Heim C, Nemeroff CB. Neurobiology of early life stress: clinical studies. *Semin Clin Neuropsychiatry* (2002) **7**(2):147–59. doi:10.1053/scnp.2002.33127
25. Luscher B, Shen Q, Sahir N. The GABAergic deficit hypothesis of major depressive disorder. *Mol Psychiatry* (2011) **16**(4):383–406. doi:10.1038/mp.2010.120
26. Wieronska J, Pałucha-Poniewiera A, Nowak G, Pilc A. Depression viewed as a GABA/glutamate imbalance in the central nervous system. In: Juruena M, editor. *Clinical, Research and Treatment Approaches to Affective Disorders*. Sao Paolo: InTech (2012). p. 235–66.
27. Castren E. Is mood chemistry? *Nat Rev Neurosci* (2005) **6**(3):241–6. doi:10.1038/nrn1629
28. Marchetti I, Koster EW, Sonuga-Barke E, Raedt R. The default mode network and recurrent depression: a neurobiological model of cognitive risk factors. *Neuropsychol Rev* (2012) **22**(3):229–51. doi:10.1007/s11065-012-9199-9
29. Sheline YI, Barch DM, Price JL, Rundle MM, Vaishnavi SN, Snyder AZ, et al. The default mode network and self-referential processes in depression. *Proc Natl Acad Sci U S A* (2009) **106**(6):1942–7. doi:10.1073/pnas.0812686106
30. Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL. A default mode of brain function. *Proc Natl Acad Sci U S A* (2001) **98**(2):676–82. doi:10.1073/pnas.98.2.676
31. Greicius MD, Krasnow B, Reiss AL, Menon V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc Natl Acad Sci U S A* (2003) **100**(1):253–8. doi:10.1073/pnas.0135058100
32. Hamilton JP, Chen MC, Gotlib IH. Neural systems approaches to understanding major depressive disorder: an intrinsic functional organization perspective. *Neurobiol Dis* (2013) **52**(0):4–11. doi:10.1016/j.nbd.2012.01.015
33. Wiebking C, de Greck M, Duncan NW, Heinzel A, Tempelmann C, Northoff G. Are emotions associated with activity during rest or interoception? An exploratory fMRI study in healthy subjects. *Neurosci Lett* (2011) **491**(1):87–92. doi:10.1016/j.neulet.2011.01.012
34. Damoiseaux JS, Rombouts SA, Barkhof F, Scheltens P, Stam CJ, Smith SM, et al. Consistent resting-state networks across healthy subjects. *Proc Natl Acad Sci U S A* (2006) **103**(37):13848–53. doi:10.1073/pnas.0601417103
35. Wei M, Qin J, Yan R, Li H, Yao Z, Lu Q. Identifying major depressive disorder using hurst exponent of resting-state brain networks. *Psychiatry Res* (2013) **214**(3):306–12. doi:10.1016/j.pscychresns.2013.09.008
36. Anderson MC, Green C. Suppressing unwanted memories by executive control. *Nature* (2001) **410**(6826):366–9. doi:10.1038/35066572
37. Seeley WW, Menon V, Schatzberg AF, Keller J, Glover GH, Kenna H, et al. Dissociable intrinsic connectivity networks for salience processing and executive control. *J Neurosci* (2007) **27**(9):2349–56. doi:10.1523/JNEUROSCI.5587-06.2007
38. Schlösser RGM, Wagner G, Koch K, Dahnke R, Reichenbach JR, Sauer H. Fronto-cingulate effective connectivity in major depression: a study with fMRI and dynamic causal modeling. *Neuroimage* (2008) **43**(3):645–55. doi:10.1016/j.neuroimage.2008.08.002
39. Mayberg HS. Targeted electrode-based modulation of neural circuits for depression. *J Clin Invest* (2009) **119**(4):717–25. doi:10.1172/JCI38454
40. Ebmeier KP, Donaghey C, Steele JD. Recent developments and current controversies in depression. *Lancet* (2006) **367**(9505):153–67. doi:10.1016/S0140-6736(06)67964-6
41. Beck AT. *Depression: Causes and Treatment*. Philadelphia, PA: University of Pensylvania Press (2006).
42. Hyler SE. APA online CME practice guideline for the treatment of patients with major depressive disorder. *J Psychiatr Pract* (2002) **8**(5):315–9. doi:10.1097/00131746-200209000-00008
43. Waite J, Easton A. *The ECT Handbook: The Third Report of the Royal College of Psychiatrists' Special Committee on ECT*. London: The Royal College of Psychiatrists (2013).
44. Kim DR, Pesiridou A, O'Reardon JP. Transcranial magnetic stimulation in the treatment of psychiatric disorders. *Curr Psychiatry Rep* (2009) **11**(6):447–52. doi:10.1007/s11920-009-0068-z
45. Linden DEJ, Habes I, Johnston SJ, Linden S, Tatineni R, Subramanian L, et al. Real-time self-regulation of emotion networks in patients with depression. *PLoS One* (2012) **7**(6):e38115. doi:10.1371/journal.pone.0038115
46. Mathews A, MacLeod C. Cognitive vulnerability to emotional disorders. *Annu Rev Clin Psychol* (2005) **1**:167–95. doi:10.1146/annurev.clinpsy.1.102803.143916
47. Li B, Piriz J, Mirrione M, Chung C, Proulx CD, Schulz D, et al. Synaptic potentiation onto habenula neurons in the learned helplessness model of depression. *Nature* (2011) **470**(7335):535–9. doi:10.1038/nature09742
48. Gotlib IH, Joormann J. Cognition and depression: current status and future directions. *Annu Rev Clin Psychol* (2010) **6**:285–312. doi:10.1146/annurev.clinpsy.121208.131305
49. Alexander GE, Crutcher MD. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends Neurosci* (1990) **13**(7):266–71. doi:10.1016/0166-2236(90)90107-L
50. Menon V. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn Sci* (2011) **15**(10):483–506. doi:10.1016/j.tics.2011.08.003
51. Bressler SL, Menon V. Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn Sci* (2010) **14**(6):277–90. doi:10.1016/j.tics.2010.04.004
52. Li R, Wang S, Zhu L, Guo J, Zeng L, Gong Q, et al. Aberrant functional connectivity of resting state networks in transient ischemic attack. *PLoS One* (2013) **8**(8):e71009. doi:10.1371/journal.pone.0071009
53. Speechley WJ, Woodward TS, Ngan ET. Failure of conflict to modulate central executive network activity associated with delusions in schizophrenia. *Front Psychiatry* (2013) **4**:113. doi:10.3389/fpsyt.2013.00113
54. Barch DM. Brain network interactions in health and disease. *Trends Cogn Sci* (2013) **17**(12):603–5. doi:10.1016/j.tics.2013.09.004
55. Palaniyappan L, Simmonite M, White TP, Liddle EB, Liddle PF. Neural primacy of the salience processing system in schizophrenia. *Neuron* (2013) **79**(4):814–28. doi:10.1016/j.neuron.2013.06.027
56. Manoliu A, Riedl V, Zherdin A, Muhlau M, Schwerthoffer D, Scherr M, et al. Aberrant dependence of default mode/central executive network interactions on anterior insular salience network activity in schizophrenia. *Schizophr Bull* (2014) **40**(2):428–37. doi:10.1093/schbul/sbt037
57. Vytal K, Hamann S. Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *J Cogn Neurosci* (2010) **22**(12):2864–85. doi:10.1162/jocn.2009.21366
58. Panksepp J. Affective neuroscience of the emotional BrainMind: evolutionary perspectives and implications for understanding depression. *Dialogues Clin Neurosci* (2010) **12**(4):533–45.
59. Ekman P. Handbook of cognition and emotion. In: Dalgleish T, Power T, editors. *Basic Emotions*. Sussex: John Wiley & Sons, Ltd (1999). p. 45–60.
60. Damasio AR, Grabowski TJ, Bechara A, Damasio H, Ponto LL, Parvizi J, et al. Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nat Neurosci* (2000) **3**(10):1049–56. doi:10.1038/79871
61. Barrett L, Wager T. The structure of emotion: evidence from neuroimaging studies. *Curr Dir Psychol Sci* (2006) **15**(2):79–83. doi:10.1111/j.0963-7214.2006.00411.x
62. Cacioppo J, Berntson G, Larsen J, Poehlmann K, Ito T. The psychophysiology of emotion. 2nd ed. In: Lewis M, Haviland-Jones R, editors. *The Handbook of Emotions*. New York, NY: Guilford Press (2000). p. 173–91.
63. Honey CJ, Kotter R, Breakspear M, Sporns O. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc Natl Acad Sci U S A* (2007) **104**(24):10240–5. doi:10.1073/pnas.0701519104
64. Zhou C, Zemanova L, Zamora G, Hilgetag CC, Kurths J. Hierarchical organization unveiled by functional connectivity in complex brain networks. *Phys Rev Lett* (2006) **97**(23):238103. doi:10.1103/PhysRevLett.97.238103
65. Deco G, Jirsa V, McIntosh AR, Sporns O, Kotter R. Key role of coupling, delay, and noise in resting brain fluctuations. *Proc Natl Acad Sci U S A* (2009) **106**(25):10302–7. doi:10.1073/pnas.0901831106
66. Leonardi N, Richiardi J, Gschwind M, Simioni S, Annoni JM, Schluep M, et al. Principal components of functional connectivity: a new approach to study dynamic brain connectivity during rest. *Neuroimage* (2013) **83**:937–50. doi:10.1016/j.neuroimage.2013.07.019
67. Allen EA, Damaraju E, Plis SM, Erhardt EB, Eichele T, Calhoun VD. Tracking whole-brain connectivity dynamics in the resting state. *Cereb Cortex* (2014) **24**(3):663–76. doi:10.1093/cercor/bhs352

68. Liu X, Duyn JH. Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proc Natl Acad Sci U S A* (2013) **110**(11):4392–7. doi:10.1073/pnas.1216856110
69. Pinotsis DA, Hansen E, Friston KJ, Jirsa VK. Anatomical connectivity and the resting state activity of large cortical networks. *Neuroimage* (2013) **65**(0):127–38. doi:10.1016/j.neuroimage.2012.10.016
70. Diaz BA, Van Der Sluis S, Moens S, Benjamins JS, Migliorati F, Stoffers D, et al. The Amsterdam resting-state questionnaire reveals multiple phenotypes of resting-state cognition. *Front Hum Neurosci* (2013) **7**:446. doi:10.3389/fnhum.2013.00446
71. Johnstone T, van Reekum CM, Urry HL, Kalin NH, Davidson RJ. Failure to regulate: counterproductive recruitment of top-down prefrontal-subcortical circuitry in major depression. *J Neurosci* (2007) **27**(33):8877–84. doi:10.1523/JNEUROSCI.2063-07.2007
72. Anand A, Li Y, Wang Y, Wu J, Gao S, Bukhari L, et al. Activity and connectivity of brain mood regulating circuit in depression: a functional magnetic resonance study. *Biol Psychiatry* (2005) **57**(10):1079–88. doi:10.1016/j.biopsych.2005.02.021
73. Sheline YI, Price JL, Yan Z, Mintun MA. Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. *Proc Natl Acad Sci U S A* (2010) **107**(24):11020–5. doi:10.1073/pnas.1000446107
74. Mitterschiffthaler MT, Kumari V, Malhi GS, Brown RG, Giampietro VP, Brammer MJ, et al. Neural response to pleasant stimuli in anhedonia: an fMRI study. *Neuroreport* (2003) **14**(2):177–82. doi:10.1097/00001756-200302100-00003
75. Keedwell PA, Andrew C, Williams SC, Brammer MJ, Phillips ML. The neural correlates of anhedonia in major depressive disorder. *Biol Psychiatry* (2005) **58**(11):843–53. doi:10.1016/j.biopsych.2005.05.019
76. Dunn RT, Kimbrell TA, Ketter TA, Frye MA, Willis MW, Luckenbaugh DA, et al. Principal components of the Beck depression inventory and regional cerebral metabolism in unipolar and bipolar depression. *Biol Psychiatry* (2002) **51**(5):387–99. doi:10.1016/S0006-3223(01)01244-6
77. Heller AS, Johnstone T, Light SN, Peterson MJ, Kolden GG, Kalin NH, et al. Relationships between changes in sustained fronto-striatal connectivity and positive affect in major depression resulting from antidepressant treatment. *Am J Psychiatry* (2013) **170**(2):197–206. doi:10.1176/appi.ajp.2012.12010014
78. Wang Y, Xu C, Cao X, Gao Q, Li J, Liu Z, et al. Effects of an antidepressant on neural correlates of emotional processing in patients with major depression. *Neurosci Lett* (2012) **527**(1):55–9. doi:10.1016/j.neulet.2012.08.034
79. Hoflich A, Baldinger P, Savli M, Lanzenberger R, Kasper S. Imaging treatment effects in depression. *Rev Neurosci* (2012) **23**(3):227–52. doi:10.1515/revneuro-2012-0038
80. MacQueen G, Born L, Steiner M. The selective serotonin reuptake inhibitor sertraline: its profile and use in psychiatric disorders. *CNS Drug Rev* (2001) **7**(1):1–24. doi:10.1111/j.1527-3458.2001.tb00188.x
81. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. *Trends Cogn Sci* (2012) **16**(1):72–80. doi:10.1016/j.tics.2011.11.018
82. Wang XJ, Krystal JH. Computational psychiatry. *Neuron* (2014) **84**(3):638–54. doi:10.1016/j.neuron.2014.10.018
83. Deco G, Kringelbach ML. Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron* (2014) **84**(5):892–905. doi:10.1016/j.neuron.2014.08.034
84. Worgotter F, Porr B. Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. *Neural Comput* (2005) **17**(2):245–319. doi:10.1162/0899766053011555
85. Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. Midbrain dopamine neurons encode decisions for future action. *Nat Neurosci* (2006) **9**(8):1057–63. doi:10.1038/nn1743
86. Schultz W. Getting formal with dopamine and reward. *Neuron* (2002) **36**(2):241–63. doi:10.1016/S0896-6273(02)00967-4
87. Deserno L, Boehme R, Heinz A, Schlagenhauf F. Reinforcement learning and dopamine in schizophrenia: dimensions of symptoms or specific features of a disease group? *Front Psychiatry* (2013) **4**:172. doi:10.3389/fpsyt.2013.00172
88. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* (1998) **393**(6684):440–2. doi:10.1038/30918
89. Newman ME. Modularity and community structure in networks. *Proc Natl Acad Sci U S A* (2006) **103**(23):8577–82. doi:10.1073/pnas.0601602103
90. Bassett DS, Bullmore ET. Human brain networks in health and disease. *Curr Opin Neurol* (2009) **22**(4):340–7. doi:10.1097/WCO.0b013e32832d93dd
91. Liu Y, Liang M, Zhou Y, He Y, Hao Y, Song M, et al. Disrupted small-world networks in schizophrenia. *Brain* (2008) **131**(Pt 4):945–61. doi:10.1093/brain/awn018
92. Zhang T, Wang J, Yang Y, Wu Q, Li B, Chen L, et al. Abnormal small-world architecture of top-down control networks in obsessive-compulsive disorder. *J Psychiatry Neurosci* (2011) **36**(1):23–31. doi:10.1503/jpn.100006
93. Wang L, Zhu C, He Y, Zang Y, Cao Q, Zhang H, et al. Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. *Hum Brain Mapp* (2009) **30**(2):638–49. doi:10.1002/hbm.20530
94. Borsboom D, Cramer AOJ, Schmittmann VD, Epskamp S, Waldorp LJ. The small world of psychopathology. *PLoS One* (2011) **6**(11):e27407. doi:10.1371/journal.pone.0027407
95. Peng D, Shi F, Shen T, Peng Z, Zhang C, Liu X, et al. Altered brain network modules induce helplessness in major depressive disorder. *J Affect Disord* (2014) **168**:21–9. doi:10.1016/j.jad.2014.05.061
96. Balaguer-Ballester E, Lapish CC, Seamans JK, Durstewitz D. Attracting dynamics of frontal cortex ensembles during memory-guided decision-making. *PLoS Comput Biol* (2011) **7**(5):e1002057. doi:10.1371/journal.pcbi.1002057
97. Niessing J, Friedrich RW. Olfactory pattern classification by discrete neuronal network states. *Nature* (2010) **465**(7294):47–52. doi:10.1038/nature08961
98. Bathellier B, Ushakova L, Rumpel S. Discrete neocortical dynamics predict behavioral categorization of sounds. *Neuron* (2012) **76**(2):435–49. doi:10.1016/j.neuron.2012.07.008
99. Samsonovich A, McNaughton BL. Path integration and cognitive mapping in a continuous attractor neural network model. *J Neurosci* (1997) **17**(15):5900–20.
100. Rolls ET, Loh M, Deco G, Winterer G. Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. *Nat Rev Neurosci* (2008) **9**(9):696–709. doi:10.1038/nrn2462
101. Rolls ET, Loh M, Deco G. An attractor hypothesis of obsessive-compulsive disorder. *Eur J Neurosci* (2008) **28**(4):782–93. doi:10.1111/j.1460-9568.2008.06379.x
102. Poldrack RA. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci* (2006) **10**(2):59–63. doi:10.1016/j.tics.2005.12.004
103. Castellanos FX, Di Martino A, Craddock RC, Mehta AD, Milham MP. Clinical applications of the functional connectome. *Neuroimage* (2013) **80**:527–40. doi:10.1016/j.neuroimage.2013.04.083
104. Deco G, Jirsa VK, McIntosh AR. Resting brains never rest: computational insights into potential cognitive architectures. *Trends Neurosci* (2013) **36**(5):268–74. doi:10.1016/j.tins.2013.03.001
105. Wilson HR, Cowan JD. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys J* (1972) **12**(1):1–24. doi:10.1016/S0006-3495(72)86068-5
106. Silver RA. Neuronal arithmetic. *Nat Rev Neurosci* (2010) **11**(7):474–89. doi:10.1038/nrn2864
107. Turrigiano G. Homeostatic synaptic plasticity: local and global mechanisms for stabilizing neuronal function. *Cold Spring Harb Perspect Biol* (2012) **4**(1):a005736. doi:10.1101/cshperspect.a005736
108. Posner J, Hellerstein DJ, Gat I, Mechling A, Klahr K, Wang Z, et al. Antidepressants normalize the default mode network in patients with dysthymia. *JAMA Psychiatry* (2013) **70**(4):373–82. doi:10.1001/jamapsychiatry.2013.455
109. Abbott CC, Lemke NT, Gopal S, Thoma RJ, Bustillo J, Calhoun VD, et al. Electroconvulsive therapy response in major depressive disorder: a pilot functional network connectivity resting state FMRI investigation. *Front Psychiatry* (2013) **4**:10. doi:10.3389/fpsyt.2013.00010
110. Messina I, Sambin M, Palmieri A, Viviani R. Neural correlates of psychotherapy in anxiety and depression: a meta-analysis. *PLoS One* (2013) **8**(9):e74657. doi:10.1371/journal.pone.0074657
111. Jedynak M, Pons AJ, Garcia-Ojalvo J. Cross-frequency transfer in a stochastically driven mesoscopic neuronal model. *Front Comput Neurosci* (2015) **9**:14. doi:10.3389/fncom.2015.00014
112. Mayberg HS, Lozano AM, Voon V, McNeely HE, Seminowicz D, Hamani C, et al. Deep brain stimulation for treatment-resistant depression. *Neuron* (2005) **45**(5):651–60. doi:10.1016/j.neuron.2005.02.014
113. Tokutsu Y, Umene-Nakano W, Shinkai T, Yoshimura R, Okamoto T, Katsuki A, et al. Follow-up study on electroconvulsive therapy in treatment-resistant

depressed patients after remission: a chart review. *Clin Psychopharmacol Neurosci* (2013) **11**(1):34–8. doi:10.9758/cpn.2013.11.1.34
114. Friedman NP, Miyake A, Young SE, Defries JC, Corley RP, Hewitt JK. Individual differences in executive functions are almost entirely genetic in origin. *J Exp Psychol Gen* (2008) **137**(2):201–25. doi:10.1037/0096-3445.137.2.201
115. Bojak I, Oostendorp TF, Reid AT, Kotter R. Connecting mean field models of neural activity to EEG and fMRI data. *Brain Topogr* (2010) **23**(2):139–49. doi:10.1007/s10548-010-0140-3
116. Gitelman DR, Penny WD, Ashburner J, Friston KJ. Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. *Neuroimage* (2003) **19**(1):200–7. doi:10.1016/S1053-8119(03)00058-2
117. Fuchs A. Beamforming and its applications to brain connectivity. In: Jirsa VK, McIntosh AR, editors. *Handbook of Brain Activity*. Berlin: Springer Verlag (2007). p. 357–78.
118. Daunizeau J, David O, Stephan KE. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *Neuroimage* (2011) **58**(2):312–22. doi:10.1016/j.neuroimage.2009.11.062
119. Friston KJ, Harrison L, Penny W. Dynamic causal modelling. *Neuroimage* (2003) **19**(4):1273–302. doi:10.1016/S1053-8119(03)00202-7
120. Schuyler B, Ollinger JM, Oakes TR, Johnstone T, Davidson RJ. Dynamic causal modeling applied to fMRI data shows high reliability. *Neuroimage* (2010) **49**(1):603–11. doi:10.1016/j.neuroimage.2009.07.015
121. Smith SM, Miller KL, Salimi-Khorshidi G, Webster M, Beckmann CF, Nichols TE, et al. Network modelling methods for FMRI. *Neuroimage* (2011) **54**(2):875–91. doi:10.1016/j.neuroimage.2010.08.063
122. Friston KJ, Dolan RJ. Computational and dynamic models in neuroimaging. *Neuroimage* (2010) **52**(3):752–65. doi:10.1016/j.neuroimage.2009.12.068
123. Valdes-Sosa PA, Roebroeck A, Daunizeau J, Friston K. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage* (2011) **58**(2):339–61. doi:10.1016/j.neuroimage.2011.03.058
124. Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Schofield TM, et al. Comparing families of dynamic causal models. *PLoS Comput Biol* (2010) **6**(3):e1000709. doi:10.1371/journal.pcbi.1000709
125. Daunizeau J, Stephan KE, Friston KJ. Stochastic dynamic causal modelling of fMRI data: should we care about neural noise? *Neuroimage* (2012) **62**(1):464–81. doi:10.1016/j.neuroimage.2012.04.061
126. Moran R, Pinotsis DA, Friston K. Neural masses and fields in dynamic causal modeling. *Front Comput Neurosci* (2013) **7**:57. doi:10.3389/fncom.2013.00057
127. Pittenger C, Duman RS. Stress, depression, and neuroplasticity: a convergence of mechanisms. *Neuropsychopharmacology* (2008) **33**(1):88–109. doi:10.1038/sj.npp.1301574
128. Johansen JP. Neuroscience: anxiety is the sum of its parts. *Nature* (2013) **496**(7444):174–5. doi:10.1038/nature12087
129. Fox MD, Snyder AZ, Vincent JL, Corbetta M, Van Essen DC, Raichle ME. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc Natl Acad Sci U S A* (2005) **102**(27):9673–8. doi:10.1073/pnas.0504136102
130. Haller S, Kopel R, Jhooti P, Haas T, Scharnowski F, Lovblad KO, et al. Dynamic reconfiguration of human brain functional networks through neurofeedback. *Neuroimage* (2013) **81**:243–52. doi:10.1016/j.neuroimage.2013.05.019
131. Koush Y, Rosa MJ, Robineau F, Heinen K, Rieger WS, Weiskopf N, et al. Connectivity-based neurofeedback: dynamic causal modeling for real-time fMRI. *Neuroimage* (2013) **81**:422–30. doi:10.1016/j.neuroimage.2013.05.010
132. Fitzgerald PB, Srithiran A, Benitez J, Daskalakis ZZ, Oxley TJ, Kulkarni J, et al. An fMRI study of prefrontal brain activation during multiple tasks in patients with major depressive disorder. *Hum Brain Mapp* (2008) **29**(4):490–501. doi:10.1002/hbm.20414
133. Fuster JM. *The Prefrontal Cortex*. New York, NY: Raven Press (1997).
134. Brocker DT, Swan BD, Turner DA, Gross RE, Tatter SB, Koop MM, et al. Improved efficacy of temporally non-regular deep brain stimulation in Parkinson's disease. *Exp Neurol* (2013) **239**:60–7. doi:10.1016/j.expneurol.2012.09.008
135. Fries P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn Sci* (2005) **9**(10):474–80. doi:10.1016/j.tics.2005.08.011
136. Fraiman D, Balenzuela P, Foss J, Chialvo DR. Ising-like dynamics in large-scale functional brain networks. *Phys Rev E Stat Nonlin Soft Matter Phys* (2009) **79**(6 Pt 1):061922. doi:10.1103/PhysRevE.79.061922
137. Thirion B, Varoquaux G, Dohmatob E, Poline JB. Which fMRI clustering gives good brain parcellations? *Front Neurosci* (2014) **8**:167. doi:10.3389/fnins.2014.00167
138. Lohmann G, Erfurth K, Muller K, Turner R. Critical comments on dynamic causal modelling. *Neuroimage* (2012) **59**(3):2322–9. doi:10.1016/j.neuroimage.2011.09.025
139. Liljestrom M, Kujala J, Jensen O, Salmelin R. Neuromagnetic localization of rhythmic activity in the human brain: a comparison of three methods. *Neuroimage* (2005) **25**(3):734–45. doi:10.1016/j.neuroimage.2004.11.034
140. Alexander B, Warner-Schmidt J, Eriksson T, Tamminga C, Arango-Lievano M, Ghose S, et al. Reversal of depressed behaviors in mice by p11 gene therapy in the nucleus accumbens. *Sci Transl Med* (2010) **2**(54):54ra76. doi:10.1126/scitranslmed.3001079

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 October 2014; accepted: 11 February 2015; published online: 26 February 2015.
*Citation: Bielczyk NZ, Buitelaar JK, Glennon JC and Tiesinga PHE (2015) Circuit to construct mapping: a mathematical tool for assisting the diagnosis and treatment in major depressive disorder. Front. Psychiatry **6**:29. doi: 10.3389/fpsyt.2015.00029*
This article was submitted to Systems Biology, a section of the journal Frontiers in Psychiatry.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on