FREE ENERGY IN PSYCHOANALYSIS AND NEUROSCIENCE

EDITED BY: Mark L. Solms, Peter Fonagy, Christoph Mathys and Jim Hopkins
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-581-2
DOI 10.3389/978-2-88945-581-2

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

FREE ENERGY IN PSYCHOANALYSIS AND NEUROSCIENCE

Topic Editors:
Mark L. Solms, University of Cape Town, South Africa
Peter Fonagy, University College London, United Kingdom
Christoph Mathys, Aarhus University, Denmark
Jim Hopkins, University College London, United Kingdom

Citation: Solms, M. L., Fonagy, P., Mathys, C., Hopkins, J., eds. (2020). Free Energy in Psychoanalysis and Neuroscience. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88945-581-2

Table of Contents

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 26 September 2017
doi: 10.3389/fpsyg.2017.01695

Hierarchical Recursive Organization and the Free Energy Principle: From Biological Self-Organization to the Psychoanalytic Mind

Patrick Connolly[1]* *and Vasi van Deventer*[2]

[1] Counselling and Psychology, Hong Kong Shue Yan University, Hong Kong, Hong Kong, [2] Psychology, University of South Africa, Pretoria, South Africa

OPEN ACCESS

Edited by:
Christoph Mathys,
Scuola Internazionale di Studi Superiori Avanzati (SISSA), Italy

Reviewed by:
Jim Hopkins,
University College London, United Kingdom
Karl Friston,
University College London, United Kingdom
Tamara Fischmann,
International Psychoanalytic University Berlin, Germany

***Correspondence:**
Patrick Connolly
patrickconnolly@live.com

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: 02 April 2017
Accepted: 14 September 2017
Published: 26 September 2017

Citation:
Connolly P and van Deventer V (2017) Hierarchical Recursive Organization and the Free Energy Principle: From Biological Self-Organization to the Psychoanalytic Mind. Front. Psychol. 8:1695. doi: 10.3389/fpsyg.2017.01695

The present paper argues that a systems theory epistemology (and particularly the notion of hierarchical recursive organization) provides the critical theoretical context within which the significance of Friston's (2010a) Free Energy Principle (FEP) for both evolution and psychoanalysis is best understood. Within this perspective, the FEP occupies a particular level of the hierarchical organization of the organism, which is the level of biological self-organization. This form of biological self-organization is in turn understood as foundational and pervasive to the higher levels of organization of the human organism that are of interest to both neuroscience as well as psychoanalysis. Consequently, central psychoanalytic claims should be restated, in order to be located in their proper place within a hierarchical recursive organization of the (situated) organism. In light of the FEP the realization of the psychoanalytic mind by the brain should be seen in terms of the evolution of different levels of systematic organization where the concepts of psychoanalysis describe a level of hierarchical recursive organization superordinate to that of biological self-organization and the FEP. The implication of this formulation is that while "psychoanalytic" mental processes are fundamentally subject to the FEP, they nonetheless also add their own principles of process over and above that of the FEP. A model found in Grobbelaar (1989) offers a recursive bottom-up description of the self-organization of the psychoanalytic ego as dependent on the organization of language (and affect), which is itself founded upon the tendency toward autopoiesis (self-making) within the organism, which is in turn described as formally similar to the FEP. Meaningful consilience between Grobbelaar's model and the hierarchical recursive description available in Friston's (2010a) theory is described. The paper concludes that the valuable contribution of the FEP to psychoanalysis underscores the necessity of reengagement with the core concepts of psychoanalytic theory, and the usefulness that a systems theory epistemology—particularly hierarchical recursive description—can have for this goal.

Keywords: psychoanalysis, neuropsychoanalysis, free energy principle, systems theory, hierarchical recursive organization

INTRODUCTION

A question that is at the heart of the neuropsychoanalytic project is the relationship between two levels of organization within the human organism, between the neurological level and the mental one. Professor Karl Friston's free energy principle (FEP henceforth) of biological self-organization has captured the imagination of many within both the neuroscientific and psychoanalytic fields as providing a very important new link in our understanding of the body-mind relationship. However, it is important to understand this development within some form of theoretical context that clarifies the correct implications that this development has within the growing science of neuropsychoanalysis, so that the importance of its role and influence is neither over- nor under-estimated.

It is the view of the present paper that a systems theory epistemology provides the critical theoretical context within which the significance of Friston's FEP is best understood. Systems theory provides the concept of a "recursive description" of organization of complex systems in the physical world, in which the FEP enters at a particular level of that hierarchical organization, which is the level of biological self-organization. This form of biological self-organization is in turn understood as foundational and pervasive to the higher levels of organization of the human organism that are of interest to both neuroscience as well as psychoanalysis.

The implication of adopting this hierarchical, recursive description of organization of the human organism is twofold. First, it implies that all levels of organization recursively superordinate to the level of biological self-organization (in this case the brain and the mind) must be constrained by the FEP. The second implication is that those recursively superordinate levels of the organism which are the brain and the mind, must also be subject to further principles of organization *not* fully explained by the FEP. Historically, the theoretical field of psychoanalysis reflected an effort to generate such superordinate principles of organization that obtain at the level of psychic organization. However, the field developed independently of neuroscience and biology, which meant that psychoanalytic theories were never adequately integrated within a hierarchical model of levels of organization of the human organism.

In this manner, it will be suggested that core constructs in psychoanalytic theory should be restated, in order to be located in their proper place within a hierarchical recursive organization of the (situated) organism. Further it is argued that a correct understanding of the place of the FEP in organizing the psyche in turn necessitates a restatement of the core Freudian concepts within a systems theory framework that can successfully integrate biological levels of organization with psychological ones.

This paper will first introduce systems theory and the notion of hierarchical recursive organization, and describe how findings in different sciences support such a notion of recursive levels of organization in nature. Next, the paper highlights a problem within the psychoanalytic literature, in which organizing principles such as the pleasure principle, have never been adequately connected to organizing principles in the nervous system or the body in general. Following this it is suggested that psychoanalytic principles of organization need to be restated within a recursive description of organization within the human organism which demonstrates its dependence on biological self-organization. The FEP is then described as a key regulatory principle of self-organization which recursively underlies psychoanalytic regulatory principles. The FEP is described here as a formalization of Maturana and Varela's (1980) concepts of autopoiesis (or self-making) and the structural coupling of the organism with its environment.

The importance of the FEP in bridging the physical material of the body and the nervous system with the level of organization of information and (Bayesian) beliefs—within the psychological domain—is clarified. This concept is then used to restate the question as to how the brain realizes the psychoanalytic mind as one of the evolution of different levels of systemic organization in which the concepts of psychoanalysis are viewed as describing a level of hierarchical recursive organization superordinate to that of biological self-organization and the FEP. The implication of this formulation is presented, which is that while "psychoanalytic" mental processes are fundamentally subject to the FEP, they nonetheless also add their own principles of process over and above that of the FEP.

The paper then presents an example of how psychoanalytic regulatory principles can be founded on biological self-organization through the model presented in Grobbelaar (1989) which offers a recursive bottom-up description of the organization of the "psychoanalytic" consciousness as dependent on the organization of language (and affect), which is itself founded upon the tendency toward autopoiesis within the organism. Meaningful consilience between Grobbelaar's model and the hierarchical recursive description available in Friston's (2010a) theory is described. The paper concludes that the valuable contribution of Friston's FEP to psychoanalysis underscores the necessity of reengagement with the core concepts of psychoanalytic theory, and the usefulness that a systems theory epistemology—particularly recursive description—can have for this goal.

GENERAL SYSTEMS THEORY AND THE SELF-ORGANIZATION OF SYSTEMS

When the field of general systems theory came to the fore toward the middle of the Twentieth century, the key purpose of this field was to offer an explanatory paradigm for how systems of various kinds came to regulate themselves, and generate their own principles of organization. While this question of self-regulation spanned a number of different fields, a very exemplary question was that of biological systems, and how biological systems appeared to regulate themselves, since their behavior is not directly regulated by their environment. This question can also be stated in terms of "bottom-up" or "top-down" forms of organization in a system, or how a system comes to develop top-down principles of self-regulation, that order the activity of lower-order levels of the system. The field of general systems theory was from the start also associated with the field

of cybernetics, which is the science of regulation or control of systems, which emerged at roughly the same time as general systems theory (Wiener, 1965; Von Bertalanffy, 1969/2009).

The answer that began to emerge from general systems theory is that those self-regulatory or top-down principles of organization of systems emerged from the activity of the lower-order elements themselves (in other words from bottom-up activity). This principle was proposed as applying to a wide variety of phenomena, including inorganic ones such as complex weather patterns that emerge from the interactions of vast numbers of air and water molecules in the atmosphere (Wiener, 1965), or patterns of convection that emerge in heated liquids (Prigogine and Stengers, 1984).

This same principle was applied to the regulation of biological systems, including complex social behavior: an example of the emergence of a patterned hierarchy of social dominance among chickens, is found in Wiener and Schadé (1965) restated vividly here in Grobbelaar (1989):

> *"… the pecking order [of chickens] which is generated through the interactions of the chickens is spontaneously generated out of the activity of pecking. So that the activity of pecking determines the pattern of dominance, which in turn determines who will peck who." (p. 137).*

Grobbelaar goes on to say that there are other factors that influence the pecking order, which is correct. However, the point being made is that the operation of the elements of a system and the interactions between them will generate a new form of organization which in turn comes to determine the activity of those constituent elements. We will see below that exactly the same circular causality emerged subsequently in physics; specifically in the context of synergetics where slow macroscopic (superordinate) modes of behavior enslave fast microscopic (subordinate) levels (Haken, 1983; Tschacher and Haken, 2007). In physics, this is known as the enslaving principle and emerges in things like the Center Manifold Theorem in dynamical systems theory (Carr, 1981).

What's important about this perspective is to note that though von Bertalanffy (together with others) sought to describe the self-regulation of systems based on bottom up processes, he also stressed the significance of aspects of holism and integration in the emergence of self-regulatory tendencies:

> *"It was the aim of classical physics eventually to resolve natural phenomena into a play of elementary units governed by 'blind' laws of nature. This was expressed in the ideal of the Laplacian spirit which, from the position and momentum of particles, can predict the state of the universe at any point in time. … In contrast to this mechanistic view, however, problems of wholeness, dynamic interaction and organization have appeared in the various branches of modern physics. … It is necessary to study not only parts and processes in isolation, but also to solve the decisive problems found in the organization and order unifying them, … Again, similar trends appeared in Psychology. While classical association psychology attempted to resolve mental phenomena into elementary units—psychological atoms as it were—such as elementary sensations and the like, gestalt psychology showed the existence and primacy of psychological wholes which are not a summation of elementary units and governed by dynamic laws." (Von Bertalanffy, 1969/2009, p. 31).*

This paragraph articulates what has become a core tenet of systems theory which is that, from the interaction of the lower order constituent elements of a system, *an entirely new form of organization emerges*, one which cannot be fully explained by the basic principles of interaction of the constituent elements, even though it emerges from their interaction (Haken and Levi, 2012). This process of emergent self-regulation is most fully described by the concept of recursive organization in systems theory, which is described next.

RECURSIVE EPISTEMOLOGY IN SYSTEMS THEORY

The concept of recursion is used in number of fields, and has slight variations in its meaning across some of these different fields. Within mathematics and computer science a recursive function is one whose term involves calling itself, with each successive application of the function referred to as an "iteration" (Shoenfield, 2001). Within the broad fields of systems theory and cybernetics, the term has also been used in different but related ways by Bateson (1978), Beer (1972), Keeney (1983), and Maturana and Varela (1980).

As described by Keeney (1983), a primary assumption of recursive organization of systems is that a system may be described as having different levels of organization of its activity. A second assumption is that the organization at higher-order levels influences the activities at lower levels of description. Keeney gives the example of how one might view a dance between two partners as recursive levels of organization. For example, the first partner in the dance may step to their right; this basic level of behavior could be considered the lowest level of organization in the current scheme. However, a higher level of organization refers to the level of interaction: the first partner steps to their right, while the second partner steps to their left. The activities at the level of behaviors are subordinated to this level of interaction. Yet a higher level of organization is at the level of choreography or pattern of interaction: the dance is a waltz. The activities at the level of behavior as well as interaction are subordinated to this pattern of choreography.

Keeney (1983) uses this formulation to describe a problematic pattern of marital interaction. The husband says he nags because the wife withdraws, while the wife says she withdraws because the husband nags. However, we might understand the behaviors (nagging and withdrawing) as being subordinated to a pattern of interaction which might be stated as withdraw, nag, withdraw, nag, withdraw, nag, crisis, reset (the pattern could equally begin with "nag" instead of "withdraw"). This systems-based formulation which indicates that behaviors within relationships are organized by stable patterns of interaction has come to have very strong empirical support over the decades-long work of John Gottman and colleagues in marital interaction patterns (Gottman et al., 2002)

Beyond these first two assumptions, a further assumption of recursive epistemology refers to the idea that the higher levels of organization of the system emerge from the activities at lower levels of organization. Similar to the earlier example of the "pecking order" of chickens emerging from the behaviors of pecking (Grobbelaar, 1989), so the pattern of marital interaction described above may actually emerge from nagging and withdrawing behaviors to begin with. However, the pattern becomes self-organizing over time, and begins to organize the nagging and withdrawing behaviors, so that they come to have a predictable pattern.

A final proposition of recursive organization is the principle that though higher levels of organization come to define the activities at lower levels, they cannot violate the principles of organization of those lower levels. In other words, the pattern of marital interaction of nagging and withdrawing that emerges between the marriage partners cannot consist of behaviors or emotions that the partners themselves are not capable of producing. The pattern of activity that emerges from the behaviors and interactions of a system's elements must lie within the parameters of possible behaviors or states of the system and its elements (Grobbelaar, 1989). In other words, while the higher-order levels of organization come to dominate the activity of lower-order elements, it cannot violate the lower-order principles of organization of those elements, nor exceed the range of potential actions of those elements.

Such a notion of emergent self-organization has received tangible support from a range of research trajectories. Hermann Haken developed a model of self-organization of coherent laser light, and how this self-organizing shift distinguishes it from non-coherent light. Haken's model, and the theoretical field it has given rise to (Synergetics) has become an established research trajectory across several disciplines, including biology (the "swarm" intelligence), computer engineering (AI studies) and molecular robotics (Haken and Levi, 2012).

In his book entitled "Reality is not what it seems," Rovelli (2016) of the Centre de Physique Theorique in Marseille, points toward the long-standing problem in physics which is the apparent fact that principles of physics which hold true at the macroscopic level of general relativity, do not hold at the microscopic level described by quantum mechanics, and vice versa. Though substantial efforts in the field of physics have attempted to bridge these two levels, no satisfying solutions have yet been found. Rovelli shows how research into loop quantum gravity has suggested that when basic subatomic quanta of gravitational fields cluster together in complex relationships, these aggregates start to interact with one another and develop novel behaviors that are unique to that level of aggregation. Rovelli shows how such aggregates can be described as occupying different levels of organization, and argues that quite profound changes occur once quanta aggregate up to the level of curving spacetime, meaning that activities at that level simply cannot be predicted by principles obtaining at subatomic levels alone, though they do emerge from activities at those levels (Rovelli, 2016).

The key implication that this view of recursive organization has for systems, is that a complex system may be understood as consisting of a number of levels of organization, each of which organizes the structure and therefore behavior tendencies of the system. These layers are hierarchical, in that the highest level has the greatest influence over the behavior of the system though it remains constrained by the lower levels, and all the levels are always operative, and not in competition with one another. A last point is that the developmental history of the system indicates that each successive layer of organization emerges from the layer below (which is why it cannot violate the principle of organization at the lower layer), and then *entrains* that subordinate level such that the subordinate regulatory principles now come to serve the superordinate ones, which now have a greater influence over the system's further behavior (Keeney, 1983; Grobbelaar, 1989).

For example if we adopt a two-level scheme which consists of the principle of natural selection (or survival of the fittest) and a second principle which is that culture exerts an organizational influence on the structural development of the person (especially the brain), we could think of the cultural organization as superordinate and the influence of natural selection as subordinate. In other words, we could understand the rise of cultural organization as emergent from organization through natural selection, due to the survival advantages bestowed by group membership and communication. However, over time and development, the influence of culture becomes self-organizing such that its influence is not fully explained by the principle of natural selection, and (as the superordinate emergent principle) it comes to have a greater influence over the regulation of the system: while our daily behaviors may (hopefully) mostly have the tendency of enhancing our survival, they are more specifically shaped by cultural information and norms. However, the principle of natural selection continues to operate: as long as there is any pattern to who survives and reproduces, evolution continues to take place. Further, we could say that the principle of natural selection becomes entrained by the influence of culture, such that our evolution comes to be influenced more and more in the direction of adaptation toward a cultural environment (a study of the cultural influence on evolution can be found in Richerson and Boyd's, 2006 work "Not by Genes Alone"). Armed with this concept of hierarchical recursive organization (referred to as HRO for the remainder of the text), we now move toward exploring a central difficulty in the historical development of psychoanalytic theory.

FREUD'S "PROJECT" AND THE SELF-REGULATION OF THE NERVOUS SYSTEM

Though Freud was not influenced by the growth of systems theory, in his posthumously published work "The Project for a Scientific Psychology" (1950), he set himself a task that was very similar to that prescribed by von Bertalanffy. In "the Project" his stated task was to demonstrate that all principles of human behavior, affect and psychical activity were determined by the interaction of neurons (the "constituent elements" of the psychical system, in his view), and the influence they exerted on one another through an energy he termed "Qn":

> *"The intention is to furnish a psychology that shall be a natural science: that is, to represent psychical processes as quantitatively determinate states of specifiable material particles, thus making those processes perspicuous and free from contradiction. Two principal ideas are involved: (1) What distinguishes activity from rest is to be regarded as Q [referring to the term 'quantity' described below], subject to the general laws of motion. (2) The neurones are to be taken as the material particles." (Freud, 1950, p. 295)*

He then proposed that the nervous system is regulated by a single organizing principle (he called it the primary principle), which is to divest itself of this energy (Qn), which he called a principle of *inertia*, though elsewhere referred to it as a principle of constancy. He tried to explain instances of the nervous system refraining from discharging energy as the result of the influence of a secondary principle, serving the interests of behaving adaptively with regard to the environment.

The important consideration here about "the Project" is the fact that he very clearly wanted to generate an entirely "bottom-up" description of the self-regulatory activity of the nervous system. Through defining different types of neurones and the types of influence (and barriers to influence) they exerted on one another, he hoped to explain the entirety of operation of psychical processes as complex as consciousness, memory and attention purely through these basic energic interactions of different types of neurons. He expressly avoided describing any process that could not be traced back to this basic interaction of different types of neurons (Connolly, 2016).

The similarity between the aims of "the Project" and the principles of the growing systems theory field of cybernetics (self-regulation) were remarked upon by Strachey in his translator's introduction to the text:

> *"It has been plausibly pointed out that in the complexities of the 'neuronal' events described here by Freud, and the principles governing them, we may see more than a hint or two at the hypotheses of information theory and cybernetics in their application to the nervous system. To take a few instances of this similarity of approach, we may note first Freud's insistence on the prime necessity for providing the machine with a 'memory'; again, there is his system of 'contact-barriers,' which enables the machine to make a suitable 'choice,' based on the memory of previous events, between alternative lines of response to an external stimulus; and, once more, there is, in Freud's account of the mechanism of perception, the introduction of the fundamental notion of feed-back as a means of correcting errors in the machine's own dealings with the environment." (Strachey in Freud, 1950, p. 292–293)*

However, Freud failed in this endeavor. Once he discovered that he could not overcome a number of internal contradictions in the system he had designed, he abandoned the project, and tried to have it suppressed, later stating:

> *"I can no longer understand the state of mind in which I hatched out the 'Project'" (Freud, 1950, p. 285)*

This failure appears to have been significant, in that from this point on, Freud appeared to begin to accept the use of top-down principles of regulation of the psyche for which he had not been able to generate a "bottom-up" explanation. This is really the start of his distinction of the "psychological" theory from neurology, in which he sought to describe principles that organized psychic life, even though he could not offer a description on how those principles emerged from its organic base:

> *"I have no inclination at all to keep the domain of the psychological floating, as it were, in the air, without any organic foundation. But I have no knowledge, neither theoretically nor therapeutically, beyond that conviction, so I have to conduct myself as if I had only the psychological before me" (Freud, 1898/1985, p. 26)*

A good example of this appears in his next major text which was "The Interpretation of Dreams" (1900/1991). In it, Freud introduces the concept of a preconscious gate, which limits access to consciousness of psychic material on the basis of whether the discharge of their energy causes pleasure or unpleasure, though no adequate physiological description of that pleasure or unpleasure is articulated in the text (Grobbelaar, 1989; Connolly, 2016).

Freud's effort in "the Project" sought to describe bottom-up processes that generated the self-regulation of the psyche and behavior, and in this respect bears similarity with the aims of general systems theory. However, Grobbelaar (1989) has argued that the failure of Freud's theorizing in this regard was not due to a lack of effort or diligence, but due to the lack of an adequate systems-based epistemology.

Referring back to the earlier quote by Von Bertalanffy (1969/2009) regarding the need to study emergent principles of holism and organization, not just elementary units moved by "blind" or mechanical laws of nature, this same point can be made with regard to Freud's project. If we agree with von Bertalanffy, we might suggest that the primary reason Freud's "project" failed was because he tried to explain the operation of the system based entirely on the basic principles of operation of the base elements (his description of specific types of neurons and the energy transfer between them). Essentially, we might say that he failed because he did not recognize the core insight of systems theory and HRO, which is that the basic energic interactions between types of neurons that he described in "the Project" should give rise to an entirely new form of (superordinate) organization, not fully explained by the basic energic interactions he described. We might agree with Freud's idea in "the Project" that the principles that organize the psyche and behavior might emerge from the more basic principles of energic interaction of neurons, but rather we should not agree that they can be fully explained by the principles governing that basic interaction.

THE EXAMPLE OF THE PLEASURE PRINCIPLE AND PSYCHIC ENERGY

To illustrate the importance of this distinction, a good example might be that of the pleasure principle in psychoanalysis, which is the tendency of the psyche to maximize pleasure and minimize unpleasure (Freud, 1911/1963). It is important to note that the pleasure principle appears to operate as a relatively fundamental principle ordering human behavior and psychic life,

and so we might think of it as an important "top-down" or superordinate form of regulatory principle. However, from the beginning of his theorizing about the pleasure principle, Freud attempted to generate a bottom-up explanation for it through basic processes of energic interactions of neurons, with his theory of psychic energy. In "the Project" (1950), he initially stated that a discharge of energy from the neurons was pleasurable, while an "accumulation" of energy was unpleasurable. However, after the difficulties met in "the Project," this link between pleasure and energic principles already began to fray in chapter 7 of "The Interpretation of Dreams" (1900/1991) where Freud suggested that discharges of energy could sometimes also be unpleasurable to the psyche and the preconscious gate somehow became the decisive process that allowed pleasurable discharge but opposed unpleasurable discharge, though as stated above, Freud could not offer a bottom-up physiological description in that text for how the preconscious gate might make that distinction. Freud (1920/1955) returned to this problem in "Beyond the Pleasure Principle" where he defined bound and unbound states of energy (cathexis) but in the same paper he questioned whether pleasure and unpleasure could be defined in terms of bound or unbound energy. In that paper he then made an interesting suggestion that pleasure may be linked to the *rate of change* of discharge, but never developed that idea further in his work (though the reader is referred to an exploration of this topic from the FEP perspective, where the intensity of emotion as well as its positive or negative valence, is linked to the rate of change of free energy, found in Joffily and Coricelli, 2013).

Thus, the failure of adequately linking the pleasure principle with energic processes may be an example of the problem defined above, in that the pleasure principle may emerge from basic energic interactions between neurons but can't be fully described by these. The same can be said for the energic theory itself. Freud had high hopes for his energic theory: in "the Project" (1950) he claimed that the tendency toward discharge was the fundamental motivation of all thought, emotion and behavior, and energic principles are recognized as a core metapsychological foundation of psychoanalysis (Rapaport and Gill, 1959). However, after his difficulties in describing energy as a neuronal physiological quantity in the project, he no longer attempted to describe it in terms of cathexis of Qn in neurons, despite continuing to use concepts of cathexis, binding and discharge for much of his career. Like the pleasure principle, energic concepts became described as "top-down" organizing principles of the nervous system that were not adequately described in terms of how they emerged from the basic interactions of the nervous system elements.

A central purpose of this paper is to demonstrate how useful the concept of HRO can be in linking these bottom-up and top-down levels, and indeed, this concept can address this problem of the emergence of organizing principles such as the pleasure principle. We could reformulate our definition of psychoanalytic principles of regulation of the psyche such as the pleasure principle (or another like it) as a recursively higher level of organization of the nervous system, that must nonetheless emerge from the basic interactions of the nerves themselves. While we might say that the pleasure principle (as formulated by Freud, 1911/1963) cannot violate the basic principles of organization of the nerves and their interaction, it also cannot be adequately modeled by those basic principles of interaction. This difference of organizational levels is proposed as the key reason for the failure of "the Project" (Freud, 1950), as well as the difficulty faced by Freud throughout his career (and by many subsequent psychoanalytic writers) to link the principles of organization of the psyche with those of the basic interactions of the nervous system. Had Freud been armed with a recursive epistemology, he would probably not have tried to write "the Project" or rather, may have taken a different approach to the material.

RECURSIVE EPISTEMOLOGY AND THE PROBLEM OF DIFFERENT PRINCIPLES OF ORGANIZATION AT PHYSICAL AND MENTAL LEVELS

Beyond this example of the pleasure principle and basic interactions of neurons, it can be stated that the underlying problem is really a deeper one which is the relationship between the principles which organize the structure of the body (including the nervous system) with those that appear to regulate the mind, and subjective experience. We could restate this particular aspect of the mind-body problem as a statement that the mind occupies a higher (superordinate) level of recursive organization in the person than the body does. However, this statement by itself doesn't add much to our understanding, beyond implying certain assumptions about the superordinate/subordinate relationships between the levels. What is needed is a more specific analysis of the principles of organization occurring at these different levels and defining a process whereby the emergence of the recursively higher level is explained.

The idea that mind and body, or mind and nervous system, occupy different levels of organization of the same system is not a new idea in psychoanalysis; a number of authors have not only expressed such a viewpoint but also attempted to reformulate some core psychoanalytic concepts from this viewpoint, notably including work by Grossman (1992), Seligman (2005) as well as Rosenblatt and Thickstun (1970, 1977, 1984). However, despite these efforts, systems theory epistemology, and recursive organization in particular, has never gained meaningful visibility in the mainstream of psychoanalytic (or neuropsychoanalytic) thinking.

However, the rapidly growing interest in the FEP may indeed demand a better understanding of these systems concepts from those members of the psychoanalytic community interested in the FEP. Friston's FEP is so important to psychoanalysis because it reflects a critical step forward to solving the problem of differing principles of organization at neural and psychological levels. It is the purpose of the present paper to demonstrate how this is so, as well as the necessity of a concept of recursive organization in order to make sense of the level of organization that Dr. Friston's work belongs to, and a correct understanding of its relationship to the level of constructs central to psychoanalysis. This is necessary not only to recognize the

powerful potential that the FEP has as a core metapsychological principle within psychoanalysis, but also to avoid overstating its role, and recognize the limitations the principle has for application within psychoanalysis as well. In order lay the groundwork to clarify this role of the FEP in the psychoanalytic scheme, we will clarify what is meant by a recursive description of the psyche, by following the indications expressed by Grobbelaar (1989).

RECURSIVE LEVELS OF ORGANIZATION IN PSYCHOANALYSIS

Grobbelaar (1989) stated that:

> *"... the view which is currently maintained by convention can be seen to constitute some sort of hierarchy with at its lowest level the inorganic domain, at the next level the organic, and finally at the highest level, the informational domain (Stoker, 1969). Although the components and their properties differ from one level to the next, the person as a system is constituted by the relations which obtain between the components at the same level as well as between components on different levels which defines the person as a unity. Furthermore it is clear that the organization at the lowest level sets the parameters for the recursive ordering of components/elements at the next level, so that the organization at the inorganic level will be reflected in a general way at the organic level, and in an even more indirect way at the informational level... Freudian theory is an attempt to identify the common human patterns at the inorganic and organic levels which determine the informational (psychological). . . . Freudian theory furthermore hypothesized that, at the inorganic level, the principles of organization which emerge from the energic interactions are the tendencies towards tension reduction and homeostasis, which are reflected at the organic level as the pleasure principle, and at the psychological level as the hallucinatory wish-fulfilment and the process Freud described as censorship." (p. 134–136)*

What Grobbelaar is attempting to do in this quotation is show how the processes at the psychological level are founded upon processes at work on the organic level of organization which are themselves founded upon processes at the inorganic level. It should be noted that it is not inevitable that these three levels should be used to describe the human as a system. Bateson (1978) suggests an infinite regress of levels, and that the observer selects the levels of description. However, besides selecting these for the purposes of convention, these three levels are significant, precisely because they appear so different in their organization: organic life appears to be so different from the inorganic matter we observe around us, and human consciousness in turn seems so markedly different from the self-regulation of most biological organisms, though that difference might be less marked than we believe in many cases.

What such a recursive description of the organization of the human organism necessitates, is a theoretical perspective that shows how regulatory principles of the psychoanalytic mind are related to the regulatory principles of biological self-organization at the organic level, which themselves should be related to regulatory principles at an inorganic level. The key proposition of this paper is that Friston's FEP (working in concert with evolution through natural selection which is discussed later) represents a fundamental principle of biological self-organization which has not only been shown to have consilience with fundamental propositions of psychoanalysis (Hopkins, 2012; Connolly, 2016), but which is shown to be founded upon regulatory principles at the inorganic level as well (Friston and Stephan, 2007; Friston, 2013).

Figure 1 presents a three-level recursive description of psychoanalytic regulatory principles that is influenced by Grobbelaar's (1989) description (and a similar model found in Connolly, 2016), which demonstrates this hierarchical relationship. A brief narrative description of the diagram would start from the bottom, and run as follows: inorganic elements (atoms and molecules), interact with one another (within the constraints of thermodynamic principles), and come to generate a new form of organization which is organic, or biological self-organization (which operates within the constraints of the FEP, itself constrained by natural selection, described later). The predictions encoded within the organization of the body (and after an evolutionary step, the nervous system) eventually come to be "aggregated" in the sense of a self-organizing generative model, understood here to be the ego, which is regulated in turn by its own principles, such as the pleasure principle or following Connolly (2016), a tendency to maintain its own organization.

The bi-directional arrows are to indicate that the influence then becomes top-down as well, so that the behavior of body (including its inorganic components) come to be regulated by psychological level components, though in a way which does not violate the principles of FE and thermodynamics which regulate the lower levels.

The place of Friston's FEP in this scheme is to formalize biological self-organization, which means it provides a constraint within which all informational exchange processes within an organism take place (Friston, 2012). As a result, the level above (which is psychological) cannot violate the FEP. However, as will be discussed next, new organizational principles emerge at this level, so that it is not fully explained by the FEP. These principles are elaborated upon in the next few sections.

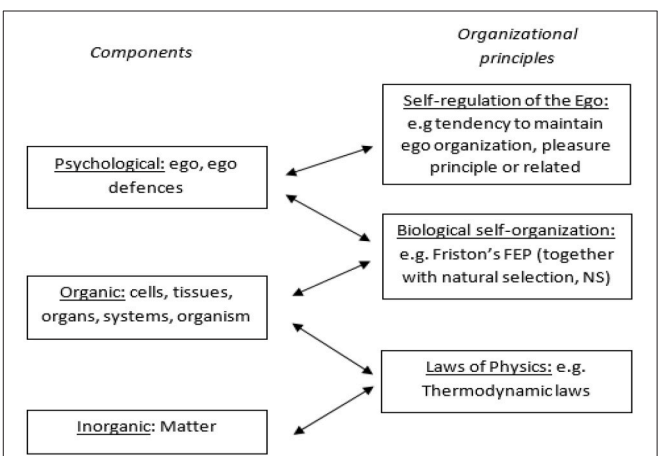

FIGURE 1 | Three levels of recursive description of organization of psychical phenomena.

FROM INORGANIC FOUNDATIONS TO BIOLOGICAL SELF-ORGANIZATION

While an engagement with levels of organization within the inorganic realm is beyond the scope of this paper, thermodynamic principles are nonetheless significant at the level of matter at which organic life occurs, and a discussion of their relationship with biological self-organization can be found in Friston and Stephan (2007). Their text states that biological systems are special in the natural universe because they appear to violate the second law of thermodynamics which is the tendency toward entropy: biological systems seem to show increasing levels of organization in their development rather than a tendency toward entropy. However, this apparent violation is just apparent—biological systems do not violate the second law but rather display a form of organization (the FEP) that is akin to the fluctuation theorems that underlie stochastic thermodynamics. These generalize the second law. As we will see later, the FEP is effectively an example of Hamilton's principle of least action. In order to develop an understanding of Friston's FEP of biological self-organization, a related verbal model (from Maturana and Varela, 1980) is offered next.

Humberto Maturana, a Chilean biologist was once asked by a student: *"'What began three thousand eight hundred million years ago so that you can say now that living systems began then?"'* (Maturana, 2002, p. 6). He realized that to answer it, he would have to identify what a living system is, and what makes it a living system. In trying to answer the question, he first made the assumption that living systems were "closed" in the sense of having an operational boundary (though being thermodynamically "open"). He then focused on the circularity of the basic metabolic reactions:

> *"… nucleic acids participate with proteins in the synthesis of proteins, and that proteins participate as enzymes with nucleic acids in the synthesis of nucleic acids, all together constituting a discrete circular dynamics. … As I was drawing a diagram of this circularity, I exclaimed 'This is it! This is the minimal expression of the circular closed dynamics of molecular productions that makes living systems discrete autonomous molecular systems.'" (Maturana, 2002, p. 7)*

From this insight, together with Varela, he coined the term "autopoiesis," which means "self-making." In other words, what constitutes a living system is two conditions. First, a boundary that creates a closed autonomous molecular structure (despite being thermodynamically open to flow of molecules in the environment). Second, a process of self-making, where a dynamic circular process takes place in which the components of the process build or maintain a structure which in turn generates the components. While this autopoietic system is closed in the sense of its organization, it is nonetheless open to the flow of molecules in the environment which participate in this self-making, and without which it would cease (Maturana and Varela, 1980; Maturana, 2002). This dependence of the autopoietic organization on the conditions in the environment led him to coin the term structural coupling.

Maturana defined living systems as structure-determined systems, in that their behavior is determined by their structure. However, that structure is occurring within a medium (or environment) and is recursively constituted from moment to moment by interactions with that medium, such that a change in one must imply a change in the other (Maturana and Varela, 1980; Maturana, 2002).

It could be said of the FEP that what it formalizes is the structural coupling of autopoietic systems. In the following section, the FEP is briefly introduced in a manner to highlight its similarity to a concept of structural coupling as a principle of biological self-organization.

FRISTON'S FEP AS STRUCTURAL COUPLING

The FEP describes how a range of biological phenomena unfolding over time can be described as the minimization of the error between the predictions afforded by a generative model of the causes of a living system's inputs and the inputs being predicted. To state it in another way, the organism's structure encodes a model of its environment (the generative model), and over time, this generative model should become a better and better predictor of the system's inputs (instigated by the environment). The minimization of this error term can be accomplished through two routes, one being the "Bayesian" updating of the generative model (to provide better error resolving predictions), the other being through the system taking an action to alter the inputs, and thereby bringing the system's sensory samples in line with predictions. While the bulk of the research that Friston and colleagues have done with the FEP have focused on describing neurophysiological processes, the principle is understood as having applications well beyond neurophysiology, and applicable to a broader biological science, including being applicable to organisms without nervous systems (Friston, 2010b, 2012).

From the above definition, a strong consilience can be seen between the FEP, and Maturana's formulation regarding structural coupling, where the structure of the organism is continuously constituted through recursive interactions with the environment. Further, Friston's FEP has shown to provide a basis for modeling Maturana's concept of autopoiesis itself. In a paper entitled "Life as we know it," Friston (2013) shows how the formation of living systems (including a simple form of autopoiesis) can be modeled using free energy minimization with only two assumptions.

The first assumption is ergodicity, which is that, over a sufficiently long time-period, the average amount of time a system spends within an accessible state is proportionate to the probability of finding the system in the state. In the language of random dynamical systems, this is equivalent to saying that the system has a random dynamical attractor; namely, a set of states that the system frequents with a high probability. The second assumption is the existence of a "Markov blanket," which corresponds to the boundary described above: Friston (2013) states that should a boundary exist where relations on one side are

conditionally independent on influences outside the boundary, this boundary will come to constitute a Markov blanket through which the internal states of a system exchange with external (environmental) states. This exchange can be formulated as an influence of the environment (external states) on the system (internal states) that is mediated by sensory components of the Markov blanket. Conversely, the influence of the system (internal states) on the environment (external states) is mediated by the active components of the Markov blanket. Note again the emergence of circular causality that has all the hallmarks of a perception and action cycle.

The existence of the Markov blanket (that separates the system from the environment in which it is immersed) implies ergodicity of the entire partition (into external states, internal states and their Markov blanket). In turn, this requires the internal states to minimize free energy as a necessary condition for the preservation of the Markov blanket. The particular aspect of free energy minimization, from the perspective of psychology and psychoanalysis, is that free energy is not just a function of states, it is a function of the probability distributions of Bayesian beliefs that are entailed (i.e., encoded) by internal states. Technically, this means the free energy is a function of a function or a functional. The key issue here is that the imperatives for self-organization are now framed in terms of probabilistic inference and beliefs. This follows from the fact that the free energy provides a proxy or bound approximation for Bayesian model evidence. Put simply, minimizing free energy necessarily maximizes the evidence for a system's (generative) model of environmental (external) states[1]. Given the circular causality above (e.g., action perception cycles), this means any system with a Markov blanket will appear to gather evidence for its own existence; which has been called self-evidencing (Hohwy, 2016)—and is closely related to early theories of self-organization such as the good regulator theorem (Conant and Ashby, 1970).

It is evident that this notion of a Markov blanket is, formally, very similar to Maturana's assumption of a closed boundary condition for the formation of a living system. From this, Friston concludes that the formation of living systems is almost inevitable in a universe that provides ergodicity and the existence of Markov blankets. Such blankets (surrounding open self-organizing processes) may be ubiquitous in the universe but the Markov blankets associated with living systems may have a particular form of hierarchical self-assembly that corresponds to the *self-making* referred to by Maturana (2002), rather than just *self-organization* which occurs throughout the natural universe.

The claims made by Maturana are not precisely the same as that made by the FEP, and are a verbal principle rather than the precise mathematical principle that is the FEP. However, Friston et al. (2015) have suggested that active inference can be viewed as a formalization of autopoiesis. The present purpose of drawing parallels between Maturana's theory and the FEP is to present an accessible verbal principle related to the FEP which allows its place within a recursive description of the mind to be perceived. From this section, it is hoped that the reader can see that what is most precisely described by the FEP is a process of biological self-organization, rather than being a purely "neurocentric" principle (Friston, 2010b), and for this reason it is presented as a fundamental constraint organizing organic systems in the current description.

THE ROLE OF EVOLUTION THROUGH NATURAL SELECTION

A key question that might be raised at this point is the role of evolution through natural selection (NS is used henceforth). Thus far, the paper has presented the FEP as the key principle of organization of biological systems, despite a wealth of evidence suggesting that the structure of organisms including the nervous system and human behavioral tendencies have been shaped by evolution through natural selection (Cartwright, 2016). For this reason, care is taken to try to explain the relationship between the FEP and NS adopted by this paper and the useful role that HRO can have in clarifying this relationship as well.

Within the biological realm, where the FEP is operative, HRO of biological systems becomes constrained by the FEP, such that the FEP may be described as a superordinate organizational principle which entrains (and constrains) HRO, resulting in further hierarchical organization being reflected in the structure of the organism, and most relevantly, the hierarchical organization of the structure of the brain, such as the layers of the mammalian cortex. The role of NS can be considered superordinate to the FEP in a similar way. In essence, once organisms are characterized by a cycle of life, reproduction and death—and provided there is some measure of environmental order influencing selection of survival and reproduction—survival of the fittest comes to act as a recursive feedback loop slowly operating at a population level, through each generational iteration. This results in a new organizational principle entraining the operation of the FEP (and thereby HRO as well), such that the organic phenotypes that exist now reflect this influence by NS. At the same time, the NS principle cannot violate the FEP, and as suggested by Hobson and Friston (2016), evolution can be understood as minimizing the free energy of specific phenotypes. As such, the different organizing principles of FE and NS are understood here as not in competition with one another, and though hierarchically arranged, all operate in organizing the organism and its behavioral tendencies[2].

Regarding the relationship of NS with the regulatory principles of psychoanalysis, connections have been made between the level of organization in psychoanalysis with that of natural selection (Hopkins, 2003, 2004, 2015; Hopkins, "Group conflict and group violence: a perspective from Freud and Darwin," forthcoming). A good example is found in Hopkins (2004; Hopkins, "Group conflict and group violence: a perspective from Freud and Darwin," forthcoming), where he describes an organizational principle emerging from natural selection, which is the tendency toward outgroup aggression, which shows how the survival advantage granted by group identification (and outgroup aggression) may underlie the

[1]My thanks to Dr Karl Friston for his helpful remarks in clarifying ergodicity, Markov blankets and free energy.

[2]My thanks to reviewer Professor Jim Hopkins for helping with the development of this reading of the relationship between the FEP and NS, by pressing my thoughts of the role of HRO and the FEP in this direction.

evolution of mechanisms of projection and introjection described by psychoanalysis.

This description of the relationship between the FEP and NS requires much more detailed discussion than is given here. However, this paper focuses on the specific role played by the FEP in the hierarchical self-organization of the organism, and particularly its relevance as a foundation of psychoanalytic principles of self-regulation. This specific and unique importance of the FEP lies in how it constrains HRO in a scale-free manner within a recursive hierarchy of levels of organization in the brain, and the nature of message passing between them, which is addressed later.

THE SCALE-FREE NATURE OF THE FEP IN THE BEHAVIOR OF BIOLOGICAL SYSTEMS

The concept of a scale-free principle is one that applies to all possible levels of scale of a phenomenon at hand (Mitchell, 2009). In other words, we would say that the principle holds no matter the scale at which you observe a phenomenon. What this means for the FEP, is that no matter at what scale you observe biological systems, the FEP should not be violated. In other words, the FEP can be observed at the level of single cells, or any components they are made of (e.g., mitochondria, dendrites), at the level of organs, systems and whole organisms (Friston, personal communication, 13th July 2015). At a neural level alone, the FEP may apply over a short time span to the activity of neurons, and over a longer time span to the reorganization of neural connections (Friston et al., 2006).

A large number of empirical findings have begun to show the variety of phenomena that can be described using the FEP formulation. These include the hierarchical deployment of cortical areas, neuromodulatory gain control and associative plasticity, receptive field effects, components of evoked cortical responses, and on a cognitive level, perceptual categorization, temporal sequencing and attention (Friston, 2010b). Such research is ongoing, and it is likely that this is just the beginning, and that there will be a substantial increase in phenomena described by the FEP, over the next years.

This apparent scale free perspective supports the notion of HRO adopted in this text, as it suggests that any hierarchically superordinate forms of organization that may develop within biological organisms, should nonetheless not violate the basic organizing principle of this organic level which is the FEP. Just as there is no action a human system can take which violates the principles of thermodynamics, so there is no action a human system can take which (viewed over a sufficiently long period) can violate the FEP. If you knew exactly what to measure and how to measure it, you could show that a person's action or thought always minimizes FE, at least when averaged over an adequate time period (technically, the average of an energy is known as a Hamiltonian action; this means that the free energy principle is a statement of Hamilton's principle of least action). Though it may be that a human system does something that appears to raise the overall level of FE in their system in the short term, the effect may be compared to dropping a ball and the principle of gravity: when it bounces and travels upwards it appears to violate the gravity principle, but over time, it will obey the principle (a similar comparison for the principle of psychic energy was found in Galatzer-Levy, 1983).

CAN ALL HUMAN BEHAVIOR BE MODELED BY THE FEP?

The above section would seem to imply that all the behavior comprising a human living system is subject to the FEP, which is indeed correct. However, while it could be claimed that the FEP as a working principle is not violated at any levels of organization of the human system (above the inorganic), this does not mean that all phenomena in a human living system are appropriately modeled using the FEP. This distinction can be displayed with an analogy.

Newton's second law of motion, force equal mass times acceleration ($F = ma$), should apply to the movement of all physical bodies in space, within particular limits in terms of mass, gravity and so on. However, if an engineer was supposed to model the complex operation of forces moving through the structure of a jet airplane as it flies through atmosphere, armed only with the model $F = ma$, it would prove to be a wildly impractical task. This would require a complete knowledge of every vector of force at work on and within the structure of the airplane at every moment, as well as perfect theoretical knowledge of how those vectors will operate from moment to moment. In other words, our engineer would need additional principles, in the form of models of "aggregate" processes such as lift, drag, stress dynamics, turbulence and others. These aggregate models involve different equations than that of $F = ma$, though none of them can violate this foundational model.

For human behaviors at the level of interest of psychoanalysis (for example actions, speech, thoughts, dreams and so on), the situation is comparable. While the previous section has described how human behavior and psychological processes at all levels cannot violate the FEP, if one were expected to model the complexity of human behavior and thought armed only with the FEP equation, it would be an equally wildly impractical task. One would need to know to know the exact state of activity of the entire nervous system (and indeed the entire body), as well as a comprehensive range of precise theoretical principles for how this state will progress from moment to moment (including how these states relate to thoughts, emotions and behaviors at the observable level). Just as in the analogy of the jet aircraft above, one would need to have a range of additional principles that model such aggregates of activity that hold at the level of interest.

The view of this paper is that the propositions of psychoanalytic theory (such as transference, repression or splitting) reflect such models at this higher level of organization, similar to lift, drag and turbulence in the engineering analogy. In the long term, the challenge is to demonstrate the relationship of these models to the foundational organization of the FEP through a recursive description of the levels of organization superordinate to the FEP, up to and beyond consciousness. This task is returned to later in this paper.

However, a question the reader might have at this point would be to ask how levels of organization of the human system that

are superordinate to the FEP, can be subject to the FEP without being sufficiently modeled by the FEP. It is hoped that these two sections have shown that it is no contradiction at all. While there is nothing a human can do that can violate the second law of thermodynamics, that law is hardly enough to model human behavior. While the organizational principle of the FEP is much closer to the level of interest that is psychoanalysis, the same limitation applies. The FEP does nonetheless retain some influence over superordinate levels through feedback loops, throughout the levels of recursion.

FRISTON'S FEP AS A MODEL OF RECURSIVE ORGANIZATION OF HIERARCHY IN THE NERVOUS SYSTEM

The diagram in **Figure 2** below demonstrates a hierarchical and recursive description of organization in the nervous system found in predictive coding formulations of Friston's FEP, whereby surprise or prediction error messages progress "upwards" through the hierarchical generative model to successively higher levels of abstraction, which respond with "downwards" predictions.

As stated earlier, Friston (2010a) has suggested that higher level predictions refer to increasing levels of abstraction. As suggested in Hobson et al. (2014):

> *"Many of the interesting insights offered by equating consciousness with the process of inference rest on the hierarchical nature of generative or virtual reality models. In hierarchical models, inference can be decomposed into multiple levels, with progressively higher or deeper levels of representational abstraction or explanation. This leads to the distinction between inferences at low levels of sensory hierarchies—that can be associated with unconscious inference in the sense of Helmholtz (1866/1962)—and at higher levels that could be associated with conscious percepts and concepts. Consider now a further hierarchical level that predicts (and selects) the particular trajectory that is enacted. This level may generate top-down predictions of proprioceptive trajectories and their visual consequences. In other words, we have moved beyond simple motor representations to a hierarchical level where expectations (neuronal activity and their associated beliefs) are quintessentially sensorimotor in nature. At this level, the multimodal nature of descending predictions (aka corollary discharge) renders the expectations amodal. Would these constitute conscious experience? One could argue that these high-level, dynamically structured beliefs are much closer to phenomenal consciousness. Furthermore, if we now equip our hierarchical model with models that distinguish between the consequences of self-made acts and the acts of others, we start to get closer to conceptual expectations of the sort that may underlie subjective consciousness."*

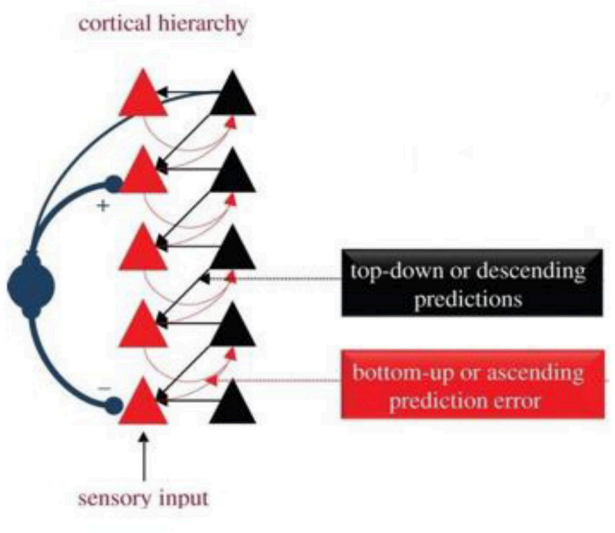

FIGURE 2 | Hierarchical organization of message passing in the brain, from Seth and Friston (2016). This illustrates the recursive message passing it implied by predictive coding formulations of free energy minimization; where in prediction errors are passed up the hierarchy and predict ions are sent down to cancel prediction errors. The blue circles and arrows denote precision that controls the influence of ascending predict ion errors. Here, precision corresponds to the inverse variability or confidence assigned to prediction errors.

This account suggests that the predictive model is organized at multiple nested layers, all of which are influenced by the FEP through this recursive feedback process. As suggested earlier in this paper, the FEP comes to constrain HRO in the development of the structure of the human organism, and in the nervous system in particular, such that each recursively higher level of organization found in the body and especially in the brain, comes to have Markov characteristics, which implies that each level has self-organizing characteristics, and tries to minimize its own free energy. Note that this hierarchical nature of the generative model induces Markov blankets between different hierarchical levels, which mediate a circular causality through recurrent message passing between levels. The existence of a Markov blankets within the brain affords the opportunity for higher levels in the brain to make inferences about lower levels (c.f., metacognition, self-modeling and consciousness). However, while all these levels are influenced by the FEP through these recursive feedback loops, it is an error to suggest that the processes at all of these levels of recursion are fully explained by the FEP. A helpful example here would be natural selection. Although the principles of natural selection can be applied to all processes of biological evolution, simply knowing these principles does not help explain the emergence of particular phenotypic traits or constructs such as convergent evolution, speciation and other emergent properties such as selection for selectability (Kauffman and Johnsen, 1991; Kauffman, 1993; Knobloch, 2001; Frank, 2012; Campbell, 2016).

THE LIMITS OF THE FEP IN MODELING CONSCIOUSNESS AND PSYCHIC EXPERIENCE

Friston (2008) has suggested that the multitude of nested levels of organization within the nervous system each have Markov properties, which implies that each level has some degree of

self-organization. This also implies that each level would require additional principles of organization (beyond the FEP) in order to be adequately described. As consciousness occupies one of the highest levels of recursion within the organization of the brain (Hobson et al., 2014), all of these subsidiary nested levels (with their own unique organizing principles) would in turn influence the organization of consciousness. As such, the very high level of complexity involved in the organization of conscious experience is practically not able to be modeled by Friston's FEP equation.

This qualification seems important, as psychologists (including psychoanalysts) are often opposed to reductionist models of conscious experience, partly because they are so abstracted from the experience itself, but often more because reductionist models often simply cannot explain the complexity of their clients' experience. In this regard, they are entirely correct, as the previous paragraph has attempted to clarify. At the same time, however, it might be said that a therapist's perception of the complexity of their client and their lived experience should not be reduced by the acceptance that the client's psychic processes cannot violate the FEP, just as it should also not be reduced by accepting that client's body cannot violate thermodynamic laws. Extraordinary levels of complexity are possible within the broader constraints of thermodynamics as well as the FEP, which in turn require detailed analysis at the level of interest as well as the proximal influences of sub- and super-ordinate levels of description.

The generations of work in psychoanalysis to document the principles which seem to influence people's conscious experience and behavior, as well as the insights gained through clinical examination and self-reflection, are understood here as attempts to generate models of the organization of the phenomena of conscious and unconscious processes. These insights (and the models they represent) cannot be abandoned in favor of a far more foundational principle which is the FEP, for much the same reason as we should not abandon the use of the abstractions of integral calculus in favor of using the simpler language of linear algebra, to follow an analogy found in Rosenblatt and Thickstun (1984).

However, like Freud (1898/1985), we cannot afford to leave these insights "floating in the air" in a completely abstract theoretical space unconnected with any organic foundation. Following a call by Grobbelaar (1989), these models of experience and behavior at the psychoanalytic level of interest need to be reformulated within a new language that demonstrates the foundations of their organization within a recursive description, which has its foundations at the biological level.

The complexity of differentiation within the physical structure of the human body is huge, and there are also already a large number of models within the biological field that predict processes within this differentiated structure. Likewise, the differentiation within the brain is also highly complex. Following Bateson (1978), there are potentially infinite levels of regress in such descriptions, and it is neither possible nor even desirable to build a complete picture of every possible level of organic and neural organization superordinate to the basic level of biological organization which is the FEP, up to the level of interest which is here psychoanalysis. Rather, it is desirable to identify some of the most significant forms of organization that are foundational to psychoanalysis, but superordinate to the FEP, which can build an intelligible bridge between the two. The description provided by Grobbelaar (1989) provides a useful example of a recursive description of this kind which may illustrate a way forward.

A RECURSIVE DESCRIPTION OF THE REPRESENTATIONAL ORGANIZATION OF CONSCIOUS EXPERIENCE

Grobbelaar (1989) offered a critique of Freud's account of the organization of consciousness (in terms of how psychic material does or does not become conscious), in that it did not offer a bottom-up recursive description:

> *"As it stands now, Freud's formulation of the process of censorship defines it as an ad hoc defensive manoeuvre by one system, the ego, against another system, the unconscious, to stop dangerous elements (dangerous to the organization of the ego) from entering the ego. One should rather formulate from the bottom to the top, that is, in a theoretical sense. One should begin by defining the inherent qualities in the lower-order elements which … make it impossible for them to be taken up in a higher order system …." (p. 142)*

He also states:

> *" … the principles determining the perception of thoughts will be inherent in the thoughts themselves. Stated differently, if the organization of the ideational domain does not allow for the representation of certain ideas, thoughts or memories, then they cannot become conscious." (pp. 139–140)*

In describing the principles inherent in thoughts which allow (or don't allow) them access to conscious, Grobbelaar (1989) refers to a comment made by Breuer in "Studies on hysteria" which refers to the notion that the quantity of affect attached to the thoughts, and the pleasure or unpleasure that that quantity of affect forms part of, determines their capacity to enter consciousness (Freud and Breuer, 1895/2004). This is related to Freud's notions that only sufficiently cathected thoughts or perceptions can enter consciousness (Freud, 1900/1991, 1950).

It can be noted at this stage that this determinant of the level of affect (or perhaps cathexis rather) has good consilience with Friston's (2010a) hierarchical description, which suggests that only information that is sufficiently surprising (or rather with sufficient gain, due to weighted precisions of surprise) can activate the predictions at the highest level of organization, which may be consciousness. Information that is insufficiently surprising is "automated" in the sense that it is sufficiently explained by predictions at lower hierarchical levels of the model, and does not elicit these higher-level predictions of consciousness (Hobson et al., 2014).

However, besides this requirement of the quantity of affect, Grobbelaar (1989) also points toward a comment made by Freud (1915/1957) in "The unconscious" where unconscious elements can only become pre-conscious through being connected with words:

> *"The system unconscious contains the thing-cathexes of the objects, the first and true object-cathexes; the system Pcs comes about by this thing presentation being hypercathected through being linked with the word presentations corresponding to it." (Freud, 1915/1957, pp. 200–201)*

Grobbelaar (1989) suggests that this formulation found in Freud (1915/1957) "The unconscious" is based on a much earlier paper on aphasia (Freud, 1891/1953) in which Freud suggests that a word-presentation is built up of a sound-image (auditory), the letter-image (visual), the motor-speech image (kinaesthetic) and the visual- and motor-writing image (visual and kinaesthetic) which become associated with one another through experience. Freud then states that the object presentation is built up in a similar way from kinaesthetic, visual and auditory experiences, and becomes linked to word presentation through associative learning, allowing for the object to reach conscious representation. The heart of Grobbelaar's argument is that this process can be described as a recursive one:

> *"This process constitutes a recursive ordering of discrete elements of experience through successive acts of integration with new elements which progressively constitute the raw sense data at higher levels of psychological functioning." (Grobbelaar, 1989, p. 147)*

Grobelaar's account of hierarchical levels of representation separated by boundaries here bears very strong similarity to a later paper by Grossman (1992) who also utilized Freud's paper on aphasia to build a similar argument. What Grossman (1992) argued is that Freud's theorizing suggests just such an underlying hierarchical model which involves discrete hierarchically defined systems with their own boundaries, where information at one level moves across boundaries as "representation" in another bounded system. At this point, it is hoped that the reader can observe the special and unique role that the FEP can play in such a description of the Freudian mind proposed by Grobbelaar (1989) and Grossman (1992). The FEP is useful here because it specifies just such a regulatory principle of the emergence of HRO within the nervous system, where hierarchically superordinate layers that emerge in the nervous system obtain self-organizing or Markov characteristics, and where message passing between layers is represented "upwards" as prediction error in progressively higher levels, and downwards as predictions (and the precisions of those predictions) to progressively lower levels. However, while the FEP provides a basis for formulating a HRO model of organization in the nervous system, it does not explain the specific operation of each of those levels and what they contribute toward the overall functioning of the system. What would be needed would be a description of the most relevant and proximal layers that most closely influence the level of interest which is that of psychoanalytic regulatory principles.

Grobbelaar does indeed go further in terms of offering such an example of a recursive description of the kind he calls for, represented diagrammatically in **Figure 3** below.

In this diagram, Grobbelaar (1989) is presenting a recursive description of the organization of conscious experience which

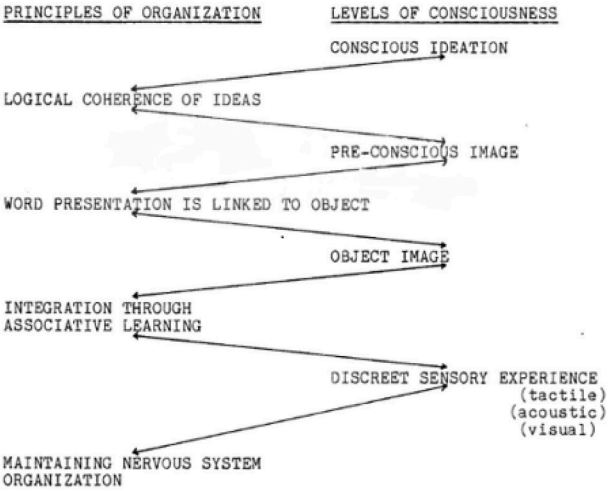

FIGURE 3 | A recursive description of consciousness organized by language, after Grobbelaar (1989).

suggests that psychic material can only become conscious when it can be represented at successively higher levels of organization. Conversely, any psychic material which is not integrated into the (recursive and hierarchical) organization cannot become conscious. Following the dependence of consciousness on language in his recursive description here, we might say that whatever experience has not yet been integrated into the linguistic organization of our brains cannot become conscious, and that (to some extent) the ego could be defined as that part of our psyche which is organized by language.

However, Grobbelaar (1989) suggests a broader understanding of what is meant by representation:

> *"The exclusivity of word-representation in allowing thoughts into consciousness should not be over-emphasized. Language should be seen as one of the most important organizing principles of experience, which acts as an entrance for experience to enter into consciousness. Its importance is linked partly to its inherent functional qualities but far more important is its quality of being used as an organizing principle. The important concept here is that experience has to be organized to have psychological meaning. It is obvious that language does not have a monopoly on this function. In the perception of music it is one's ability to perceive the rhythm and melodious organization which seems to be (a) independent of language (b) improved with repeated exposure (c) dependent on a different symbolic notation. The perception of visual pattern seems also to be dependent on an organizing principle other than language. It does, however, rely on language to an unknown extent where the visual pattern is an object which is also represented in our language … all these principles act as determinants of pattern discrimination. Without discriminating the pattern … there can be no awareness of the object." (pp. 148–149)*

A last comment is made here about the lowest level of the recursive description found in **Figure 3** above. It is important to note that this level of ordering refers to the maintenance of

organization of the nervous system. In his thesis, Grobbelaar (1989) argued that this organizing tendency of language was itself recursively constituted from the tendency toward autopoiesis, or self-making in the organism. In this way he hoped to demonstrate how the psychological ordering of conscious experience was itself founded upon the autopoietic ordering of the organism. By these means, he sought to describe the ego as self-organizing. Though Grobbelaar did not have access to Friston's (2010a) FEP, there is nonetheless good consilience between his view and a formulation of the recursive ordering of information within hierarchical layers such as that found in the FEP.

This last point is also important to understanding the powerful role that interoceptive influence plays on the activity of the mind, throughout the feedback loops described here. It assumed here that interoceptive information must enter the hierarchical organization of the brain at a relatively lower level of organization than that described by Grobbelaar (1989) above. However, a central proposition of psychoanalysis is that the homeostatic requirements of the body in the form of interoceptive information are a fundamental driver of affect and motivation. Due to the genetic endowment of the brain, as well as the primary place of interoceptive information in the early life of the organism (Hobson et al., 2014)—especially *in utero*—and the narrow parameters within which internal organs usually remain, the precisions associated with interoceptive information are very high. Therefore, a large proportion of the surprise present in the nervous system emerges from interoceptive input, particularly when activating prototype emotions, and so, when high enough, can progress through successive feedback loops up every layer structure that reflects the organizational hierarchy described here, so that even a conscious stream of logically ordered thoughts can be constrained by nagging perceptions of hunger and thoughts about what to do about it. However, it is acknowledged that this important topic of the role of interoceptive input in HRO and the FEP requires additional attention beyond that given here.

Grobbelaar's (1989) work has identified two subordinate levels or organization that are critical to conscious experience. He has chosen the organization of entry to preconsciousness as the level of interest, and generated a recursive description of the dependence of conscious organization on that of language (or patterned representation more broadly), though he also signaled the importance of affect without developing it much further in his text. The organization of affect as a foundational principle for consciousness is not explored in this article, though the reader is referred to a paper by Hopkins (2016), entitled "Free energy and virtual reality in neuroscience and neuropsychoanalysis: a complexity theory of dreaming and mental disorder," where he demonstrates how the interaction between mental states organized by conflicting emotions—attempting to minimize their respective free energy—underlie a process of conscious experience that is strongly consilient with that described by psychoanalytic theory.

The preceding sections have hopefully lain the groundwork for a key conclusion expressed here, which is that the ego must be understood as self-organizing (Grobbelaar, 1989), and that the specific nature of that self-organizing process is itself emergent from the FEP (Connolly, 2016). The ego is understood here as an associative structure occupying the higher levels of organization of the generative model, that comes to influence lower levels of the hierarchy. As such, it develops Markov characteristics that mean it (the ego) must be viewed as effectively self-organizing—and potentially self-evidencing as described by Hohwy (2016). Psychoanalysis is proposed here as being essentially that science of the self-organization of the ego that describes its relative inertia and resistance to change, while also describing the unique principles of organization that operate at this level.

CONCLUSION

The aim of Grobbelaar's (1989) argument was to challenge the notion of a top down ordering process of organizing consciousness (such as the preconscious gate in Freud, 1900/1991), that could not be shown to emerge from a bottom up process, and to call for a reformulation of Freudian concepts that makes use of systemic principles such as HRO. In turn, the thrust of the present paper is to show how such a recursive description is precisely what is needed to correctly recognize the influence of biological self-organization (in the form of the FEP) on processes related to conscious experience that are of central interest to psychoanalysis. Equally, the paper has also tried to demonstrate the limitations of the FEP in fully explaining higher levels of organization within the person, and that the self-organizing nature of the ego demands distinct models, which are what psychoanalytic concepts can be understood as offering. It is hoped that the paper offers a compelling argument in this regard. Future work should re-examine the key theoretical constructs of psychoanalysis in order to offer a recursive description of their dependence on lower levels of organization in the brain, within the constraints of Friston's FEP.

AUTHOR CONTRIBUTIONS

PC provided the key theoretical ideas and wrote the paper, and is responsible for submission. The paper would be the first related publication of a Ph.D. thesis completed in 2016. VvD was supervisor of the Ph.D. thesis.

REFERENCES

Bateson, G. (1978). "The birth of a double bind," in *Beyond the Double Bind: Communication and Family Systems, Theories and Techniques with Schizophrenics*, ed M. M. Berger (New York, NY: Brunner Mazel), 53.

Beer, S. (1972). *Brain of the Firm*. Allen Lane: The Penguin Press.

Campbell, J. O. (2016). Universal darwinism as a process of Bayesian inference. *Front. Syst. Neurosci.* 10:49. doi: 10.3389/fnsys.2016.00049

Carr, J. (1981). *Applications of Centre Manifold Theory*. Berlin: Springer-Verlag.

Cartwright, J. (2016). *Evolution and Human Behavior: Darwinian Perspectives on the Human Condition, 3rd Edn*. London: Palgrave.

Conant, R. C., and Ashby, W. R. (1970). "Every Good Regulator of a system must be a model of that system." *Int. J. Syst. Sci.* 1, 89–97. doi: 10.1080/00207727008920220

Connolly, J. P. (2016). *Principles of Organization of Psychic Energy Within Psychoanalysis: a Systems Theory Perspective*. Unpublished doctoral thesis. University of South Africa, Pretoria.

Frank, S. A. (2012). Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. *J. Evol. Biol.* 25, 2377–2396. doi: 10.1111/jeb.12010

Freud, S. (1950). "The project for a scientific psychology," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud,* Vol. 1 (1886-1899): *Pre-Psycho-Analytic Publications and Unpublished Drafts,* ed J. Strachey (London: Hogarth Press), 281–399.

Freud, S. (1891/1953). *On Aphasia: A Critical Study*. ed E. Stengel, New York, NY: International Universities Press. (Original work published 1891).

Freud, S. (1920/1955). *Beyond the Pleasure principle. The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. 18 (1920-1922): *Beyond the Pleasure Principle, Group Psychology and Other Works,* London: Hogarth Press (Original work published in 1920), 1–64.

Freud, S. (1915/1957). "The unconscious," in *The Standard Edition of the Complete Works of Sigmund Freud,* Vol. 14 *(1914-1916): On the History of the Psycho-Analytic Movement, Papers on Metapsychology and Other Works,* ed J. Strachey (London: Hogarth), 159–215. (Original work published 1898).

Freud, S. (1911/1963). "Formulations regarding the two principles of mental functioning," in *The Collected papers of Sigmund Freud*, Vol. 6: *General Psychological Theory: Papers on Metapsychology,* ed P. Rieff (New York, NY: Collier Books), 21–28. (Original work published in 1911).

Freud, S. (1898/1985). "Letter to Fliess, September 22, 1898," in *The Complete Letters of Sigmund Freud to Wilhelm Fliess 1887-1904,* ed J. M. Masson (Cambridge: Belknap Press), 326–327. (Original work published 1898).

Freud, S. (1900/1991). *The Interpretation of Dreams*. Harmondsworth: Penguin. (Original work published in 1900).

Freud, S., and Breuer, J. (1895/2004). *Studies in Hysteria*. New York, NY: Penguin Books. (Original work published in 1895).

Friston, K. J. (2008). Hierarchical models in the brain. *PLoS Comput. Biol.* 4:e1000211. doi: 10.1371/journal.pcbi.1000211

Friston, K. J. (2010a). A free energy principle for the brain. *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J. (2010b). Is the free energy principle neurocentric? *Nat. Rev. Neurosci.* 11:605. doi: 10.1038/nrn2787-c2

Friston, K. J. (2012). A free energy principle for biological systems. *Entropy* 14, 2100–2121. doi: 10.3390/e14112100

Friston, K. J. (2013). Life as we know it. *J. R. Soc. Int.* 10:20130475. doi: 10.1098/rsif.2013.0475

Friston, K. J., Kilner, J., and Harrison, L. (2006). A free energy principle of the brain. *J. Physiol.* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Friston, K. J., Levin, M., Sengupta, B., and Pezzulo, G. (2015). Knowing one's place: a free-energy approach to pattern regulation. *J. R. Soc. Int.* 12:20141383. doi: 10.1098/rsif.2014.1383

Friston, K. J., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458. doi: 10.1007/s11229-007-9237-y

Galatzer-Levy, R. M. (1983). Perspective on the regulatory principles of mental functioning. *Psychoanal. Contemp. Thought* 6, 255–289.

Gottman, J. M., Murray, J. D., Swanson, C. C., Tyson, R., and Swanson, K. R. (2002). *The Mathematics of Marriage: Dynamic Nonlinear Models.* Cambridge: MIT Press.

Grobbelaar, P. W. (1989). Freud and Systems Theory: an Exploratory Statement. Johannesburg: Rand Afrikaans University. (Unpublished doctoral dissertation).

Grossman, W. I. (1992). Hierarchies, boundaries and representation in a Freudian model of mental organization. *J. Am. Psychoanal. Assoc.* 40, 27–62. doi: 10.1177/000306519204000102

Haken, H. (1983). *Synergetics: An Introduction. Non-Equilibrium Phase Transition and Self-Selforganization in Physics, Chemistry and Biology.* Berlin: Springer Verlag.

Haken, H., and Levi, P. (2012). *Synergetic Agents: from Multi-Robot Systems to Molecular Robotics*. Weinheim: Wiley.

Hobson, J. A., and Friston, K. J. (2016). A response to our theatre critics. *J. Conscious. Stud.* 23, 245–254.

Helmholtz, H. (1866/1962). "Concerning the perceptions in general," in *Treatise on Physiological Optics, 3rd Edn* (New York, NY: Dover).

Hobson, J. A., Hong, C. C., and Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Front. Psychol. Cogn. Sci.* 5:1133. doi: 10.3389/fpsyg.2014.01133

Hohwy, J. (2016). The Self-Evidencing Brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Hopkins, J. (2003). "Emotion, evolution and conflict," in *Psychoanalytic Knowledge,* eds M. C. Chung and C. Feltham (London: Palgrave), 132–156.

Hopkins, J. (2004). "Conscience and conflict: Darwin, Freud, and the origins of human aggression," in *Emotion, Evolution and Rationality,* eds D. Evans and P. Cruse (New York, NY: Oxford University Press), 225–248.

Hopkins, J. (2012). "Psychoanalysis, representation and neuroscience: the Freudian unconscious and the Bayesian brain," in *From the Couch to the La: Psychoanalysis, Neuroscience and Cognitive Psychology in Dialogue,* eds A. Fotopoulu, D. Pfaff, and M. Conway (Oxford: Oxford University Press), 230–265.

Hopkins, J. (2015). "The significance of consilience: psychoanalysis, attachment, neuroscience and evolution," in *Psychoanalysis and Philosophy of Mind: Unconscious Mentality in the 21st Century*, eds S. Boag, L. A. W. Brakel, and V. Talvitie (London: Karnac), 47–137.

Hopkins, J. (2016). Free energy and virtual reality in neuroscience and neuropsychoanalysis: a complexity theory of dreaming and mental disorder. *Front. Psychol. Cogn. Sci.* 7:922. doi: 10.3389/fpsyg.2016.00922

Joffily, M., and Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Comput. Biol.* 9:e1003094. doi: 10.1371/journal.pcbi.1003094

Kauffman, S. (1993). *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford: Oxford University Press.

Kauffman, S. A., and Johnsen, S. (1991). Coevolution to the edge of chaos: coupled fitness landscapes, poised states, and coevolutionary avalanches. *J. Theor. Biol.* 149, 467–505. doi: 10.1016/S0022-5193(05)80094-3

Keeney, B. (1983). *Aesthetics of Change.* New York, NY: Guilford

Knobloch, F. (2001). Altruism and the hypothesis of meta-selection in human evolution. *J. Am. Acad. Psychoanal.* 29, 339–354. doi: 10.1521/jaap.29.2.339.17264

Maturana, H. R. (2002). Autopoiesis, structural coupling and cognition: a history of these and other notions in the biology of cognition. *Cybern. Hum. Knowing.* 9, 5–34.

Maturana, H. R., and Varela, F. J. (1980). *Autopoiesis and Cognition.* London: D Reidel.

Mitchell, M. (2009). *Complexity: A Guided Tour*. New York, NY: Oxford University Press.

Prigogine, I., and Stengers, I. (1984). *Order Out of Chaos*. New York, NY: Bantam.

Rapaport, D., and Gill, M. M. (1959). The points of view and assumptions of metapsychology. *Int. J. Psychoanal.* 40, 153–162.

Richerson, P. J., and Boyd, R. (2006). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago, IL: University of Chicago Press.

Rosenblatt, A. D., and Thickstun, J. T. (1970). A study of the concept of psychic energy. *Int. J. Psychoanal.* 51, 265–278.

Rosenblatt, A. D., and Thickstun, J. T. (1977). Energy, information, and motivation: a revision of psychoanalytic theory. *J. Am. Psychoanal. Assoc.* 25, 537–558. doi: 10.1177/000306517702500302

Rosenblatt, A. D., and Thickstun, J. T. (1984). The psychoanalytic process: a systems and information processing model. *Psychoanal. Enq.* 4, 59–86. doi: 10.1080/07351698409533531

Rovelli, C. (2016). *Reality Is Not What It Seems: The Journey to Quantum Gravity*. London: Allen Lane.

Seligman, S. (2005). Dynamic systems theories as a metaframework for psychoanalysis. *Psychoanal. Dialogues* 15, 285–319. doi: 10.1080/10481881509348832

Seth, A. K., and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philos. Trans. R. Soc. B* 371:20160007. doi: 10.1098/rstb.2016.0007

Shoenfield, J. R. (2001). *Recursion Theory: Lecture Notes in Logic I*. Natick: A.K. Peters.

Stoker, M. G. P. (1969). "Regulating systems in cell culture," in *Ciba Foundation Symposium-Homeostatic Regulators*, eds G. E. W. Wolstenholme and J. Knight (London: J & A Churchill), 264–275. doi: 10.1002/9780470719695.ch16

Tschacher, W., and Haken, H. (2007). Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. *New Ideas Psychol.* 25, 1–15. doi: 10.1016/j.newideapsych.2006.09.002

Von Bertalanffy, L. (1969/2009). *General Systems Theory: Foundations, Development, Applications*. New York, NY: George Braziller. (Original work published in 1969).

Wiener, N. (1965). *Cybernetics: Or Control and Communication in the Animal and the Machine, 2nd Edn*. Cambridge: MIT Press.

Wiener, N., and Schadé, J. P. (eds.). (1965). *Progress in Biocybernetics*, Vol. 2. Philadelphia, PA: Elsevier Publishing Company.

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 Psychology

CONCEPTUAL ANALYSIS
published: 19 July 2018
doi: 10.3389/fpsyg.2018.01264

Expected Free Energy Formalizes Conflict Underlying Defense in Freudian Psychoanalysis

*Patrick Connolly**

Department of Counselling and Psychology, Hong Kong Shue Yan University, Hong Kong, Hong Kong

Freud's core interest in the psyche was the dynamic unconscious: that part of the psyche which is unconscious due to conflict (Freud, 1923/1961). Over the course of his career, Freud variously described conflict as an opposition to the discharge of activation (Freud, 1950), opposition to psychic activity due to the release of unpleasure (Freud, 1990/1991), opposition between the primary principle and the reality principle (Freud, 1911/1963), structural conflict between id, ego, and superego (Freud, 1923/1961), and ambivalence (Freud, 1912/1963). Besides this difficulty of the shifting description of conflict, an underlying question remained the specific shared terrain in which emotions, thoughts, intentions or wishes could come into conflict with one another (the neuronal homolog of conflict), and most especially how they may exist as quantities in opposition within that terrain. Friston's free-energy principle (FEP henceforth) connected to the work of Friston (Friston et al., 2006; Friston, 2010) has provided the potential for a powerful unifying theory in psychology, neuroscience, and related fields that has been shown to have tremendous consilience with psychoanalytic concepts (Hopkins, 2012). Hopkins (2016), drawing on a formulation by Hobson et al. (2014), suggests that conflict may be potentially quantifiable as free energy from a FEP perspective. More recently, work by Friston et al. (2017a) has framed the selection of action as a gradient descent of expected free energy under different policies of action. From this perspective, the article describes how conflict could potentially be formalized as a situation where opposing action policies have similar expected free energy, for example between actions driven by competing basic prototype emotion systems as described by Panksepp (1998). This conflict state may be avoided in the future through updating the relative precision of a particular set of prior beliefs about outcomes: this has the result of tending to favor one of the policies of action over others in future instances, a situation analogous to defense. Through acting as a constraint on the further development of the person, the defensive operation can become entrenched, and resistant to alteration. The implications that this formalization has for psychoanalysis is explored.

Keywords: free energy principle, conflict, psychoanalysis, defense, systems theory, neuropsychoanalysis

OPEN ACCESS

Edited by:
Christoph Mathys,
Scuola Internazionale Superiore di Studi Avanzati (SISSA), Italy

Reviewed by:
Tobias Nolte,
University College London, United Kingdom
Karl Friston,
University College London, United Kingdom

***Correspondence:**
Patrick Connolly
patrickconnolly@live.com

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *04 December 2017*
Accepted: *29 June 2018*
Published: *19 July 2018*

Citation:
Connolly P (2018) Expected Free Energy Formalizes Conflict Underlying Defense in Freudian Psychoanalysis. Front. Psychol. 9:1264. doi: 10.3389/fpsyg.2018.01264

INTRODUCTION

The free-energy principle (FEP henceforth) connected to the work of Friston (Friston et al., 2006; Friston, 2010) has provided the potential for a powerful unifying theory in psychology, neuroscience and related fields that has been shown to have tremendous consilience with psychoanalytic concepts (Hopkins, 2012), and may well have tremendous potential as a unifying metapsychological principle in psychoanalysis as well (Connolly, 2016). A recent paper by Hopkins (2016), drawing on a formulation by Hobson et al. (2014) has suggested how the free energy principle provides a basis for formalizing emotional conflict as complexity which places a demand (or "affective load" following Levin and Nielsen, 2009) on the capacity of the underlying generative model that predicts sensory experience to minimize that emotional complexity[1].

Hopkins' formulation provides the seeds of a formal description of conflict in the psychoanalytic sense. He describes conflict as irresolvable complexity in the form of a complex set of simultaneous emotions that each separately motivate behavioral plans that are in conflict with one another. Hopkins (2016) draws on the example of attachment related trauma in the "strange situation" paradigm in which the disparate emotions felt by the child when their mother leaves the room - and more especially when she returns—lead to behavioral trajectories that are fundamentally in conflict with one another (e.g., anger and fear). Since there is no single action that the child can take that would simultaneously achieve predicted satisfaction for all of these conflicting trajectories, there may be persistent emotional complexity, which is Hopkins' account of trauma in this perspective.

Most importantly for this paper is how the description of conflict in his paper is founded upon the concept of free energy, and specifically the idea that distinct neural systems in the brain motivate competing plans of action which are expected to have a high cost in terms of free energy for the alternate system. In short, the Free Energy Principle (FEP) perspective suggests that a person's decision of which policy of action to follow is determined by a computation of which policy is predicted to reduce the physiological free energy (or information "surprise") the most. From this perspective psychoanalytic conflict is presented as the state where different potential policies of action have a similar level of expected free energy, creating a subjectively unpleasant state of uncertainty of what to do. However, the present formulation of conflict and defense also necessitates a metapsychological revision of the assumptions underlying the core concepts of conflict, defense and possibly repression in line with a systems theory epistemology as spelled out in Grobbelaar (1989), which is addressed in the present article as well.

The first section will briefly describe the psychoanalytic concept of conflict, and its role in shaping defensive behavior and stable personality configurations in the person. The key problem of neurological correlates (particularly quantitatively framed correlates) is presented, including the failed explanation of psychic energy. Following this, an account of conflict from a statistical free-energy principle (henceforth FEP) perspective is explored, particularly under "expected" free energy. This formalization also suggests a route through which conflict is resolved by alteration of the relative precisions of the (beliefs about) opposing action policies. This alteration of precisions presents a means of formalizing defense, which becomes entrenched over time as a constraint on the future development of the generative model. This forms the basis for exploring the inertia of the generative model in terms of opposing the installation of certain action policies in specified situations, and also as a basis for understanding resistance in psychoanalysis as well. The implications of this particular formalization of conflict for psychoanalytic theory and practice is explored.

CONFLICT IN PSYCHOANALYSIS

Freud's core interest in the psyche and behavior was in that part of the person that was influenced by conflict. While he noted that there were large sections of the psyche which were not necessarily involved in conflict, his stated interest was in the dynamic unconscious, which is that element of the psyche which is unconscious due to conflict (Freud, 1923/1961). Later writers such as Hartmann (1964) sought to explore the conflict-free elements of the ego, and broaden the scope of analysis. But within Freud's description of the psyche, conflict plays a central role in defining behavior, emotional and psychical experience, and personality. However, Freud's conceptions of the nature of conflict and the terrain in which it took place evolved throughout the course of his work.

Beginning with "The Project for a Scientific Psychology" (written in the late 1890s and published posthumously), Freud (1950) first outlined conflict in terms of an energy present in the nervous system (represented by a quantity he called "Qn" that "cathected" or was contained within, the neurons). He then described what he called the primary principle of the nervous system which was to discharge that activation, usually through the motor apparatus and motor activity. His first formulation of conflict was a principle that operated in opposition to the first; he called this opposing principle the secondary principle, which is the demand for discharge of activation to be inhibited, delayed, and modified in order to result in adaptive behavior. He suggested that this opposition took the form of what he termed "lateral cathexes" (discharge through laterally branching neurons) which drained the main channel of its activation toward discharge, and resulted in the activation being channeled around within a subsystem of the brain (the ψ-system which was a forerunner of the ego) in a way similar to liquid in a system of interconnecting pipes (Holt, 1962). In this first formulation conflict evidently takes place on the terrain of energies distributed in the nervous system. Due to some difficulties he encountered in developing this concept (which are described later in this paper), this conceptual centrality of energy as a zone of conflict became partly displaced by the experiences of pleasure and unpleasure, where the release of unpleasure results in opposition

[1]The emotional complexity that is minimized here formally refers to the Kullback–Leibler divergence between posterior beliefs about policies or courses of action, relative to prior beliefs, which will be explained later in this paper.

and suppression of mental activity that causes such unpleasurable discharge (Freud, 1990/1991), though the neurophysiological basis of this pleasure and unpleasure was never adequately described in his work. Later, Freud (1911/1963) reformulated the primary process as the pleasure-unpleasure principle, which is the tendency of the psyche to tend toward activity that produces pleasure and avoids unpleasure. In this formulation, the conflict lay between the pleasure principle and a secondary principle he now termed the reality principle, which was the need for psychic activity to generate adaptive states and behaviors by opposing the pleasure-principle. While this is formally similar to the primary and secondary process defined in "the Project," he had then moved away from formulating these principles either in neurological or in purely energic terms, though he often tried to relate the pleasure and unpleasure to energic terms in statements made throughout his work (Connolly, 2016).

With the so-called "structural" shift toward Freud's (1923/1961) tripartite model of the psyche (the familiar id, ego, and super-ego model), rather than defining conflict in terms of the distribution and opposition of energies in the nervous system, conflict was rather stated in terms of the struggle between psychic structures (or systems): the push of the id toward satisfaction, the punitive response of the super-ego, and the ego which binds these dynamic forces. The resultant behavior, emotion or psychic experience was understood as a compromise between these forces, at times expressed in energic terms, and at others as pleasure-unpleasure and the demands of reality. This defensive compromise (which protects the person from super-ego anxiety, as well as pressure for discharge from the id) is the operation of the ego (Freud, 1923/1961). Given time, and a relatively stable environment, these compromise operations gain stability, and form the recognizable characteristics of the personality.

Beyond these descriptions of conflict, Freud also focussed on the key problem of ambivalence in the human condition, and most especially in people suffering from neurosis. In "The Dynamics of the Transference," Freud (1912/1963) explored how analysis of neurotic symptoms often gave rise to powerful ambivalence (and resistance) in the transference, and suggested that the distorted defensive behaviors in such cases often reflected the difficult compromise between the powerful ambivalent emotions and their associated motives. Typical responses to the "strange situation" described in Ainsworth and Bell (1970) and Howe (2011), may reflect such behavioral compromises between these conflicting emotional demands (Hopkins, 2016).

THE PROBLEMS OF NEUROPHYSIOLOGICAL CORRELATES AND QUANTITATIVE EXPRESSION

Besides the problem of varying definitions described above, Freud's various formulations of conflict have faced more fundamental difficulties. The two key problems referred to in this paper are the problem of neurophysiological correlates of conflict, and the related problem of the quantitative terrain of conflict.

In terms of neurophysiological correlates of conflict, there is a sizeable literature outside of psychoanalysis that has attempted to discover correlates for conflict, though conflict has various definitions (and operationalization) in this literature, including decisional conflict (between equivalent alternatives), cognitive conflict (conflict between values, actions, or beliefs), informational conflict (receiving information containing contradictions), and the task of sustaining conflicting plans within consciousness, amongst others (Gray et al., 2013; Pushkarskaya et al., 2015). However, these operationalizations of conflict are not very similar to conflict in the Freudian sense, as given in the definitions at the outset of this paper, as they are mostly conscious and not apparently related to repression in any way.

Berlin and Montgomery (2017) have very recently reviewed the existing literature on neurophysiological correlates of conflict in the psychoanalytic sense. While their chapter draws together a number of interesting studies with findings that seem to have implications for conflict, few of these studies focus explicitly on conflict itself and its key neural mechanism; many of the findings relevant to conflict from this literature are more specifically about repression, suppression and dissociation, though some do have clear implications for conflict. Relevant work from this literature is from Shevrin et al. (1996) and Shevrin et al. (2013). In their approach, conflict words have been generated from transcripts of interviews with subjects, and presented subliminally and supraliminally. Measured responses were in terms of alpha power (combined amplitude and frequency measures from EEG) representing inhibitory responses toward words relevant to the conscious symptom (related to phobias in the studies). Findings appeared to show a link between unconsciously perceived conflict words acting as a prime for a greater alpha power inhibitory response toward to conflict symptom words. This supports the idea that unconscious conflicts are related to symptoms (such as phobia), and that they involve inhibitory responses relevant to symptoms, which was not found for conscious conflict words in their study.

This distinction between conscious and unconscious conflict is supported by other studies. While the role of the anterior cingulate cortex has been demonstrated in tasks involving detection and processing of conflicts related to emotion and autobiographical material (Schmeing et al., 2013), work by Dehaene et al. (2003) found that activation in the anterior cingulate cortex, which often accompanies conscious conflict monitoring tasks, was absent in subliminal conflicts. Interestingly, Anderson et al. (2004) demonstrated that conscious suppression tasks can become automated in the sense of no longer engaging conscious attention or control, after time and repetition.

More studies focus explicitly on repression than on conflict, and after a review of this field, Anderson and Hanslmayr (2014) suggest that these findings point toward the role of the lateral prefrontal cortex (PFC) in mediating inhibitory control processes, usually interacting with subcortical structures including the hippocampus and other structures encoding memories. A related set of findings (about inhibition of emotion) exist for dissociative mechanisms, such as for Depersonalization

Disorder where the right dorsolateral PFC increases attention while the left PFC inhibits the amygdala and other limbic structures. This seems related to findings for both dissociative disorders as well as hypnotic states, in which prefrontal executive structures are found to interfere in voluntary and automatic processes (Berlin and Montgomery, 2017). Another interesting finding from dissociative processes is the finding of impaired connectivity in brain areas (Krystal et al., 1998). When set alongside the findings of impaired connectivity in psychosis (Schmidt et al., 2015) together with the relative lack of activation of conflict-related brain areas (including dorsolateral prefrontal cortex) in fMRI data from participants with clinical high risk for psychosis (Colibazzi, 2016, July), these findings seem to point toward the role that connectivity must play in conflict, in the sense that a minimum level of connectivity must be in place for conflict to take the form as understood in Freud's work.

While these findings have implications for psychoanalytic conflict, they do not clarify a specific mechanism for conflict that is distinct from a mechanism for repression or dissociation. In part, this lack may be due to inherent difficulties in studying conflict empirically, in the sense that the specific triggers of conflict are unique to each person. Studies focusing on conflict may use transcripts from interviews to generate conflict-related stimuli for use in the research, such as in Shevrin et al. (2013) above. Kessler et al. (2017) similarly used participant generated lists of positive and negative life events, followed by individual psychodynamic interviews based on operationalized psychodynamic diagnosis to create a list of cue sentences, used in free association and subsequent recall tasks. The resources needed for such individualized methods have undoubtedly slowed the field down. Further, the purpose of repression (and perhaps dissociation) is to avoid or reduce conflict, and so it may be problematic to measure conflict when successful repression (or dissociation) is taking place.

A second and related problem with the psychoanalytic notion of conflict is that any explanatory theory of this conflict must not only specify a shared domain or terrain between conflicting psychic processes, but also a quantitative expression of those processes such that the outcome can be understood as the difference between these quantities. While it may be correct to say that brain regions such as the prefrontal cortex, anterior cingulate cortex, limbic system, and hippocampus may be the terrain or domain of conflict (perhaps in terms of competition for neural resources or differential activation), this does not yet clarify the specific quantitative expression of the conflict. Horowitz (1977) suggested that the concept of conflict in psychoanalysis must involve a quantitative expression of some kind:

> *"…the concept of conflict, deriving from the therapeutic method, has been central to all psychoanalytic clinical theory, whether the locale of that conflict was with the environment or was intrapsychic.* The dynamic and economic metapsychological viewpoints grew out of the clinical data of conflict [emphasis in original]. *How such a conflict concept would look without "quantitative" assumptions underlying it is unclear. It may be that a conflict concept would be untenable without those quantitative assumptions. In any event, no set of critiques of the economic hypotheses of analysis has presented a cogent set of alternatives in providing the underpinning for the dynamic viewpoints" (Horowitz, 1977, p. 563).*

We should agree with Horowitz that conflict is untenable without quantity. For conflict to take place, not only should two phenomena exist within a shared terrain in which they can interact, but they should also be able to exert some form of influence upon one another, which is meaningless if not theoretically quantifiable (Swanson, 1977).

THE FAILED SOLUTION OF PSYCHIC ENERGY

Freud's key attempt to provide this quantitative account of conflict lay in his theories of psychic energy, and the principle of inertia which was proposed as regulating those energies. This is most pronounced with regard to the original formulation of conflict (described at the outset of this paper) which is conflict between the primary and secondary processes as defined in "The Project for a Scientific Psychology" (Freud, 1950). As described above, Freud proposed this conflict as a contest between quantities of Qn in the nerves: the result was either that the energy was retained in the ψ-system or progressed toward motor discharge, or some compromise of the two. The result was essentially determined by the levels of the quantities at play.

However, this explanation of cathexes of energy within neurons failed. The energic theory has been widely critiqued by a number of authors, and a detailed review of this debate that unfolded over several decades can be found in Connolly (2016). The most common critique has been the lack of any sound empirical evidence from brain science of the energic processes as described in "The Project" (Basch, 1976; Swanson, 1977; Zepf, 2010), and what we now know about the nervous system which is that action potentials vary in terms of frequency, but not in terms of intensity or strength (Pribram and Gill, 1976). Rapaport (1960) also outlined a familiar argument that the energic processes can't be observed directly in the clinical situation. However, the key failure that Freud himself was aware of, and which led him to eventually abandon "The Project" was irresolvable internal contradiction in the proposed model. This was because his description of the higher functions of the psyche (e.g., consciousness, memory, attention, and others) relied on a linear progression of stimulus energy from the sense organs and sensory stimulus, through the ψ-system, the system of conscious experience (or ω-system) and on to motor discharge. However, once Freud tried show how this progression of energy through these systems actually produced the phenomenology of attention, consciousness, and memory, he was forced to add constructs, revise, and rework, until he eventually radically changed the entire structure in the final pages of the collected document, and never elaborated further on this change, but instead tried to suppress the text after that.

In later work, while Freud appeared to back away from further theorizing about the quantities of Qn, he retained concepts of energy and cathexis, and in "The Interpretation of Dreams," Freud (1900/1991) viewed conflict in terms of stimulus energy moving "forward" through the psychic apparatus (toward motor discharge) being opposed by inhibition from the preconscious gate to prevent unpleasurable discharge. At the same time a regressive movement of excitation backwards through the

apparatus took place (usually in sleep or hallucination), "powered" in a sense both through inhibition, as well as by a "pull" of powerful sensory memories (Connolly, 2016). Regarding this latter text, it is important to note that while Freud was still talking about a contest between theoretically quantifiable amounts of energy, he had moved further away from specifying the physiological expression of the energy, and therefore further away from specifying the physiological terrain in which conflict between these energies could take place. Though Freud (1920/1955) developed his ideas about psychic energy further in "Beyond the Pleasure Principle," he never escaped internal contradictions of his energic theory, nor came closer to clarifying its physiological substrate (Basch, 1976; Zepf, 2010), despite continuing to use concepts of cathexis, discharge, and libido throughout his career (Holt, 1962; Connolly, 2016).

Despite the failure of the energic theory to provide a working quantitative account of conflict within psychoanalysis, the problem has remained as a troubled foundation of the field until recently.

CONFLICT WITHIN A FREE ENERGY PRINCIPLE (FEP) PERSPECTIVE

Friston's (2010) free energy principle has become of rapidly growing interest to psychoanalysis due to significant formal similarities between FEP and several assumptions within psychoanalysis, including the primary principle of mental functioning (Carhart-Harris and Friston, 2010), unconsciousness and motivation (Hopkins, 2012), emotional complexity in attachment (Hopkins, 2015), wish fulfillment within dreaming (Hopkins, 2016), and the energic theory within psychoanalysis (Connolly, 2016). The FEP also has the potential to offer a quantitative basis for a formulation of conflict as well (Hopkins, 2016), which could solve the problem of a quantitative expression of energy and conflict as well as its neurophysiological substrate or terrain.

Essentially, the FEP proposes that the physical structure of all self-sustaining and adaptive creatures encodes a model of the sensory inputs emerging from their environment. The FEP then states that living systems must then, either implicitly or explicitly, minimize their variational free energy. What is meant by variational free energy here is a quantity of informational surprise or prediction error, which is the difference between the sensory states predicted by the model, and those that are actually received. This leads to living systems avoiding surprising or highly improbable states. This is consistent with a principle from physics known as Hamilton's principle of least action (determining the path of lowest value), cast in terms of information theory. Mathematically, negative surprise is the same as (log) Bayesian model evidence (Friston, 2009). This means that creatures (or people) that minimize their free energy also maximize the evidence for their model of the world, or in other words, are self-evidencing (Hohwy, 2016). Importantly for the present argument, free energy can be minimized (and model evidence maximized) by one of two routes: either by an updating of Bayesian beliefs encoded by the generative model of the organism, or by taking an action which alters the sensory inputs of the organism in such a way that the surprise (or prediction error) is reduced.

A core significance of the FEP for psychoanalytic theory, is that it offers a potentially quantifiable formalization of Freud's concept of psychic energy. However, FE is not a physical energy, but an information theoretic concept; it does not quantify a thermodynamic energy, but rather quantifies a form of information present in the system: in this case the biological organism (Friston, 2010). However, the FEP can still play a very similar role in psychoanalytic theory to that played by the energic theory (as a formalization of the core motivator and organizational principle of activity in the person and their mind), though it may necessitate the incorporation of a systems theory epistemology to adequately do so (Connolly, 2016; Connolly and van Deventer, 2017). Most importantly for the present paper, it provides a basis for understanding how psychic processes can be quantitatively expressed.

However, as indicated earlier, beyond this requirement of a quantitative expression of conflict, there is the requirement of a shared terrain in which these quantities can come to "oppose" one another. A FEP perspective supplies this formalization of the shared terrain, though noting both the scale-free nature of the FEP in the physiological organization of the organism, and also the complexity and differentiation within the organism. Essentially, we might think of the overall organism as being constituted of a massive complexity of subsystems, that extend from sub-cellular components or organelles (e.g., dendrites or mitochondria), through cells (e.g., neurons or others), tissues, organs, systems and even higher levels of recursion. The important insight here is that each of these subsystems obey the FEP, each tries to minimize its free energy. In other words, all structures in the organism that have an identifiable boundary condition (a Markov blanket) minimize their free energy. Further, the total free energy present in the whole organism is understood as the sum of the FE present in each of the constituent subsystems (Friston et al., 2015a). This has the implication that while changes in the activity of the whole system (i.e., organism) reduce the overall FE of the system, they may at the same time, *raise* the FE of specific subsystems. An example might be the organism's response to muscles enduring sustained strain; while the organism's overall FE might be lowered by the activation of the sympathetic nervous system which brings needed oxygen and nutrients to (and removes waste products from) the strained muscle tissue, the tissues of the heart itself are pushed further from equilibrium, and may experience a relative increase in FE which is attempting (though initially failing) to drive the overall system in the opposite direction, to reduce blood flow. This formulation now provides us a basis for understanding how a theoretically quantifiable form of "conflict" (in a broad sense) can take place in the terrain of information exchange between subsystems within the biological organism.

However, this definition of conflict in the organism is broader than that implied by psychoanalysis, as this form of conflict is ubiquitous to every level of scale in the organism, whereas we might say that psychoanalytic conflict occurs at a particular level of organization in the organism. We may even find a range of

similarly broad conflict phenomena at a psychological level of interest in the organism which still do not precisely equate with the psychoanalytic concept of conflict. There are a number of such potential examples.

With regard to perception, Hopkins (2012) describes a formal similarity of this form of conflict with an artificially induced binocular rivalry paradigm, where the right and left eyes are given different objects in their field of view (e.g., a face and a house). The result is that perception oscillates between seeing a house for some time, then a face for some time, and back to a house, and so on. The neural structures that encode a house image are in competition with those that encode a face image to activate a dominant perceptual inference. Should the house image become the first dominant inference, then the sensory stimuli that are associated with the face persist as surprise (prediction error); the persistent surprise feeds upwards to the higher levels again which shifts the dominant inference to that of the face, where the stimuli from the house now feed upwards as prediction error, and so on. Hopkins (2012) suggests a consilience between this process and that of the psychoanalytic unconscious where the dominating inference renders the conflicting stimuli unconscious (usually on a more enduring basis), though the surprise associated with the suppressed stimuli still motivates automatized unconscious behaviors, though they cannot become conscious inference. This scheme is easily extended to explain the repression of sexual excitement for example, where interoceptive stimuli emerging from sexual excitement remain suppressed by a more dominant inference that doesn't integrate the stimuli, which nonetheless can activate automatized behaviors. This mechanism of the unconscious is returned to later in the paper.

Another example of this type of conflict in inference at a "psychological" level of interest (close to, but perhaps not the same as psychoanalytic conflict), refers to the perception of emotion, or the occurrence of feeling two apparently contradictory emotions at the same time, which we might encapsulate in the statement "I don't know what to feel, nervous or excited." Examples of such conflicting emotion inference from experimental science might include the studies of misattribution of arousal (Dutton and Aron, 1974; White et al., 1981) where distinct sources of arousal may nonetheless activate a dominant inference, for example where arousal due to the effort and nervousness from crossing a bridge appears to increase perceptions of attractiveness of a research confederate. This highlights a potentially important aspect of active inference in exchange with the world—and one's body. Namely, one has to infer the causes of all sorts of sensations; including proprioceptive and interoceptive (i.e., motor and autonomic) signals. In other words, we have to find explanations that account for all our sensations and select the most plausible hypothesis that best explains them. This means that interoceptive inference about the state of my body contextualizes exteroceptive sensory cues concerning "where I am" and "what I am doing." This means that sensations of autonomic arousal have to be explained (away); thereby leading to the hypothesis or explanation that "I am currently in a particular emotional state."

While the researchers in the above attribution studies have not explicitly connected their experimental findings either with psychoanalytic conflict literature or with active inference, they may nonetheless demonstrate such competition for awareness between conflicting emotional signals that may potentially be explained from a FEP perspective in a similar way as the binocular rivalry findings are explained in Hopkins (2012), in that one inference tends to dominate at a time. As suggested earlier however, while these examples occur at the level of interest in psychology more broadly, they may not reflect examples at a psychoanalytic level of explanation, as they primarily reflect conflicts in perception. From the earliest phases of Freud's work, psychoanalytic conflict has been linked to the inhibition of *action*, typically through inhibition of the flow of energy toward discharge through the motor apparatus (Freud, 1950; Breuer and Freud, 1985/2004). Expected free energy offers a formal description of action selection (Friston et al., 2017a) that offers potential for such a formalization of psychoanalytic conflict.

EXPECTED FREE ENERGY AND SELECTION OF ACTION

To understand psychoanalytic conflict—from the point of view of FE minimization—it is necessary to consider a slightly nuanced aspect of the FEP; namely, active inference and planning of action, or *expected* free energy. Expected free energy refers to the predicted level of free energy after a course of action is taken. A course of action is referred to here as a policy. A priori, the probability of selecting a particular policy decreases with the free energy expected under that policy. To refer to another example found in Hopkins (2012), a person who is experiencing surprise in the sense of interoceptive signals of dehydration or thirst, may seek a glass of water, as that course of action will have the lowest expected free energy of the various possible courses of action. In this case a generative model specifies the expected free energy following alternative courses of actions (i.e., policies) and the policy that leads to the least surprising outcomes is selected (i.e., "my thirst will be quenched").

To understand the nature of expected free energy, one can decompose it in various ways. For the purposes of the current argument, one can think of expected free energy as comprising epistemic and pragmatic parts (Friston et al., 2015b). The epistemic part tries to resolve uncertainty by taking actions with high information gain—that resolve ambiguity about the state of the world (e.g., Kapur, 2003; Itti and Baldi, 2009; Mirza et al., 2016). The pragmatic part simply reflects the prior beliefs (e.g., about drinking water) or preferences ingrained in a generative model through prior experience (or perhaps epigenetics). Friston et al. (2017a) suggest that any organism that has prior beliefs about its behavior must believe it will minimize expected free energy or, more simply, resolve uncertainty under prior beliefs about what will happen to it.

Friston et al. (2017a) suggest that it works as follows: where sensory inputs generate surprise at lower levels of a hierarchy of perceptual inference, they trigger potential action plans at

higher levels. Those action plans can be evaluated in terms of the expected free energy, informed by expectations encoded by a hierarchical generative model. This is especially interesting if one considers organisms (particularly human beings) that entertain different outcomes under different choices (or "policies" of action) at the same time. This allows for selecting actions that have the smallest expected free energy (Friston et al., 2015b). The researchers contend that this approach resolves the difficult problem of selection of action into an "easy" inference problem. This selection of action policies is usually expressed as a softmax function where the probability of an action is equal to the exponential of negative expected free energy (normalized so that the sum of probabilities is one; Friston et al., 2017a) and is represented in **Figure 1** below.

The authors demonstrate how this proposed process is neuronally plausible:

> *"This reflects a process theory which associates the expected probability of a state with the probability of a neuron (or population of neurons) firing and the logarithm of this probability with postsynaptic membrane potential. In this approach, post-synaptic depolarization caused by afferent input can be interpreted as free energy gradients (or state prediction errors) that are linear mixtures of firing rates in other neurons (or populations). These prediction errors drive changes in membrane potential and subsequent firing rates (Friston et al., 2017a)."* (Connolly, 2017).

For those less familiar with the FEP, this process theory provides a concrete understanding of what the quantity of free energy is (in terms of the nervous system at least), which is the level of influence that the activity of neurons or populations of neurons have on the rest of the system.

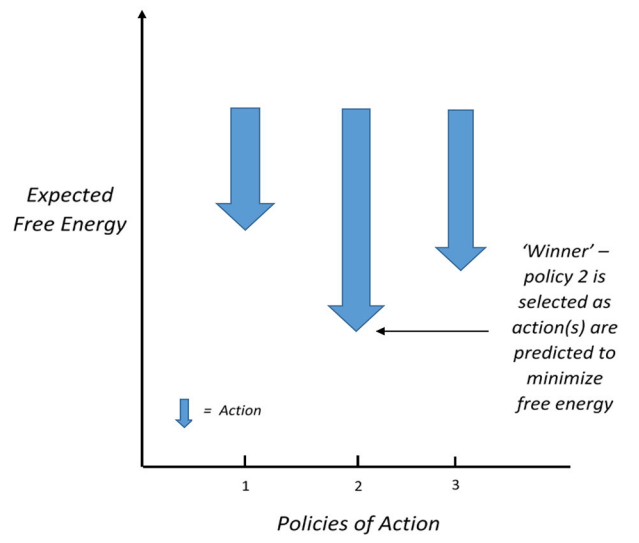

FIGURE 1 | According to Friston et al. (2017a), the nervous system calculates the expected free energy of actions under different policies of action, and those with the lowest expected free energy gain dominance (after Connolly, 2017, with permission from the copyright holder).

Besides suggesting the neuronal plausibility of this approach to action selection, Friston et al. (2017a) propose a potential functional anatomy of the process as follows:

> *"Sensory evidence is accumulated to optimize expectations about the current state of the world, which are constrained by expectations of past and future states. This corresponds to state estimation under each policy the agent entertains. The quality of each policy is evaluated in the ventral prefrontal cortex, possibly in combination with ventral striatum (van der Meer et al., 2012), in terms of its expected free energy. This evaluation and the ensuing policy selection rest on expectations about future states. Note that the explicit encoding of future states lends this scheme the ability to plan and explore. After the free energy of each policy has been evaluated, it is used to predict the subsequent hidden state through Bayesian model averaging (over policies). This enables an action to be selected that is most likely to realize the predicted outcome. Once an action has been selected, it generates a new observation, and the cycle begins again (p. 19)."*

While the authors' proposed functional anatomy is not yet supported with specific empirical proof, it is nonetheless consistent with what is generally accepted about the functional anatomy of the brain, and is presented by the authors as a possible anatomy rather than a proposed model. Its purpose is to offer support for the proposed formulation of action selection through expected free energy by describing how it might be reflected in the functioning of the brain.

Recent work on canonical microcircuits (Bastos et al., 2012) have also supported the idea that the layers of cortical columns (including functional separation of higher and lower levels of neurons and interneurons) show this form of hierarchical organization of neurons which is able to sustain the computations involved in estimating expected free energy of this kind. In this work, afferent projections from lower-order areas feedforward prediction errors which excite expectancies encoded by populations of neurons connected at that higher level; these offer inhibitory feedback connections, through stimulating inhibitory interneurons in the lower-order layers. This computational architecture allows for higher-order expectations that strongly increase free energy to more strongly inhibit lower-order stimuli that give rise to them, allowing for the phenomenology described above, which involve policies with high expected free energy to be inhibited in favor of policies which reduce the expected free energy.

The central importance that this formulation has for the present paper, is that it offers a great opportunity to provide a formal description of conflict from a FEP perspective. Conflict, from a FEP perspective can be formalized as follows: if every action has roughly the same expected free energy there is no clear winner—and the probability or belief distribution over alternative ways forward becomes uncertain; in other words, beliefs about what I am doing have a low precision. This is the mathematical homolog of conflict; namely, a loss of precision or confidence in what to do next, and is represented in **Figure 2** below.

With this model or formalism in mind, we can now see how difficult it must be for a person who entertains different actions

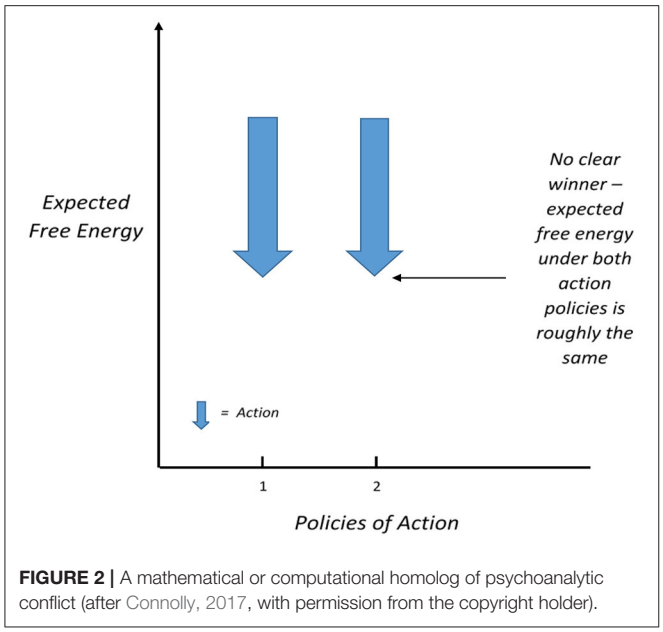

FIGURE 2 | A mathematical or computational homolog of psychoanalytic conflict (after Connolly, 2017, with permission from the copyright holder).

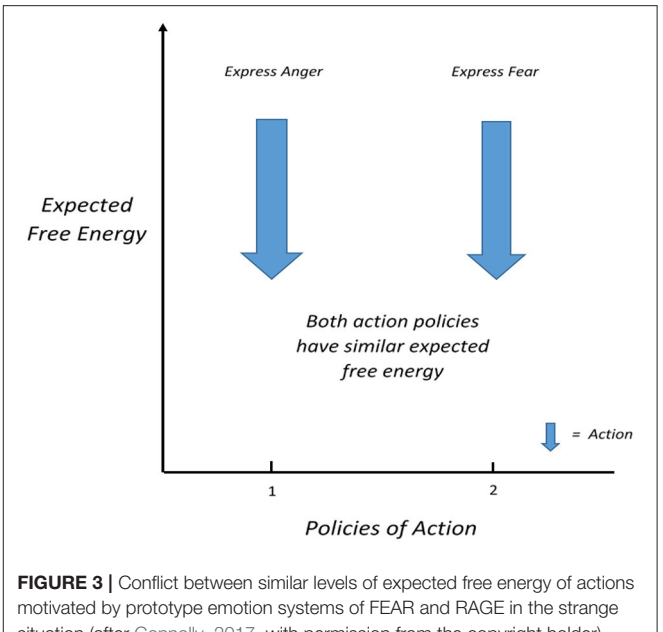

FIGURE 3 | Conflict between similar levels of expected free energy of actions motivated by prototype emotion systems of FEAR and RAGE in the strange situation (after Connolly, 2017, with permission from the copyright holder).

or policies that each only satisfy one of a number of precise prior beliefs. This is the form of irreducible uncertainty posed by conflict problems, and is consilient with the description of conflict within the psychoanalytic literature reviewed earlier.

CONFLICT IN THE STRANGE SITUATION

We may now apply this scheme to the formulation outlined in Hopkins (2016). The prototype emotion systems described by Panksepp (1998) can be understood as functional subsystems of the nervous system, ones that even have visually identifiable boundary conditions, in some respects. As described in Hopkins (2015) the strange situation simultaneously gives rise to activation of a number of these prototype emotions systems. Following the formulation in Hopkins (2012), each of these will give rise to a (best-guess) belief that a particular behavior will satisfy demand from the prototype emotion system and reduce the surprise associated with it. So, while it may be that striking the mother may satisfy the RAGE system, it is also likely that this will increase the chance of losing her, which would increase the free energy (surprise) associated with the FEAR system. In terms of active inference, what is being said here is that the particular structure of the RAGE and FEAR neural systems encodes particular prior beliefs about outcomes that can be realized by different courses of action. The conundrum here is that all available courses of action lead to violation of prior beliefs (i.e., an increase in surprise or free energy) in at least one dimension (i.e., the prototype emotions; Panksepp et al., 1984). This situation (represented in **Figure 3**) may make it difficult to reduce the FE of both of these systems, leading to persistent distress (trauma, in Hopkins' view), such as that found in the "insecurely" attached pattern of response. In this way, the insecurely attached child in the strange situation may cycle between feelings of fear and rage, much as perception cycles in the binocular rivalry paradigm described in Hopkins (2012), where the stimuli related to the currently non-dominant inference persist as surprise (i.e., unresolved prediction error) that pushes the alternative inference into dominance, and back again, and so on.

It is important to note that the example drawn from Hopkins (2016) here is an example of an application of the formal definition of conflict offered in this paper which is the situation of competing policies of action with relatively similar free energy. Theoretically, other subsystems of the brain, other than the prototype emotion systems could demonstrate this form of conflict, provided they are motivating competing action plans, and produce meaningful increases in free energy. Nonetheless, this example underlines the importance of the prototype emotion systems delineated by Panksepp (1998), and their central significance in a psychoanalytically-informed description of human consciousness and behavior as described in Solms and Turnbull (2003). Not only are these systems potent sources of free energy in the organization of the brain, but they also create the conditions for significant conflict in action selection, and explain the intensity of agitation and emotional distress caused by the persistent FE from activation not resolved by the currently dominant policy (or state estimation from predicted actions; Hopkins, 2016).

A detailed computational model of the interaction between caregiver and child that simulates the emergence of an attachment pattern in the child has recently been published by Cittern et al. (2018). In their model a Bayesian approach based on active inference (based on the FEP) is deployed within a game theoretical framework where a child agent has three available actions, "seek," "guarded seek," or "avoid." In return, a caregiving agent may be "highly responsive," "inconsistently responsive," and consistently "unresponsive," expressed in terms of "attend"

or "ignore" behavior. This model simulates a situation in which the interoceptive states of the child following "attend" or "ignore" behavior may steadily result in one of the typical organized attachment patterns of "secure," "avoidant," or "ambivalent." This simulation provides some support for the key propositions of attachment theory which is that the pattern of caregiving behavior shapes the subsequent attachment pattern of the child (Bowlby, 1969, 1973). Where affective communication errors (ACEs or cues that are misleading with regard to the subsequent behavior) are added to the model in the form of an exteroceptive cue before the "attend" or "ignore" behavior, they add further explanatory value. High levels of ACEs before inconsistent responding produced an ambivalent attachment pattern, while inconsistent cues before consistently distressing responding produced a disorganized attachment model. This is in line with research which has shown that affective communication errors are associated with both ambivalent and disorganized attachment patterns (Bronfman et al., 1999; Safyer, 2013). The observation that these models produce responses that are in line with what is expected from previous theory and research on attachment, is offered as strong support for a FEP based computational model of the child's responses to a caregiver (Cittern et al., 2018). With regard to the formulation being offered in this paper, this work supports the grounding of a process of conflict within a FEP-based model, that may be organized by interactional process as suggested by Hopkins (2016).

The present formulation of conflict also provides the basis for generating a reformulation of a psychoanalytic explanation for the development of an unconscious due to defense, through a description of how the brain processes and overcomes this conflict state through an alteration of the relative "precisions" of the demands associated with these emotion systems, described next.

DEFENSE AS ALTERED PRECISIONS

Referring back to the description of the expected free energy of action policies in Friston et al. (2017a), prior beliefs about outcomes (that underwrite pragmatic value) could themselves be inferred. This offers an insight into how conflict could be overcome or resolved: namely by assigning greater precision to a particular set of prior beliefs about outcomes to resolve ambiguity in situations of conflict. The term "precision" here refers to the range of variation allowable within incoming information; higher precision means that even minor variations in stimulus values may generate error, whereas lower precision means that incoming information can vary a lot more before generating surprise. We might use an analogy found in Peterfreund and Schwartz (1971) about a thermostat: if a thermostat only allows for a variation of two degrees on either side of the optimum temperature before activating an air conditioner, it will be activated far more easily or often than a thermostat that allows variation of four degrees on either side. The narrower range of variation allowed by the thermostat (two degrees) is similar to a situation of higher precision, where more minor variations can generate strong error.

In the above example, precision can be regarded as the sensitivity of posterior beliefs to some form of evidence; in other words, the confidence we ascribe to evidence. Exactly the same interpretation applies to the precision of beliefs about discrete outcomes; say, for example a number of competing or conflicting policies. In economics, the precision corresponds to the sensitivity parameter of a softmax function; also known as inverse temperature. In short, a precise belief distribution means that there is one clear winner and we are confident about some state of affairs. In what follows, we will start by considering the precision of beliefs about courses of action; i.e., "what am I doing"—and then drill down to the prior beliefs about outcomes that underwrite policy or action selection. The balancing of different prior preferences depends upon the precision of these preferences, emphasizing one sort of outcome over another.

Applying this to psychoanalytic conflict again, by assigning a higher precision to one of the conflicting alternatives, ambiguity or uncertainty (about policies) can be resolved because one course of action reduces expected free energy more potently than all alternative choices (there is now a clear winner, in the economics sense described above). Intuitively, this is essentially the same as assigning a greater "importance" to minimizing the surprise of one of the subsystems as opposed to the others, through altering the precision of prior preferences about outcomes encoded by these subsystems. Referring back to the previous example of prototype emotions, this would be equivalent to assigning greater importance to satisfying either the RAGE or FEAR prototype emotions, thus tipping the balance of expected free energy in favor of one or the other.

THE FORMATION OF AN UNCONSCIOUS

Significantly, the consequence of this change, is that the other non-dominant action plan is now no longer determining the action plan that becomes represented in conscious experience and is acted upon. However, the prediction error associated with this now non-dominant alternative is not entirely removed either—this persisting error may potentially be reflected in the apparently "unconscious" agitation or intensity that appears to accompany conflict in the clinical situation, even when the person is only aware of the dominant inference regarding their own mental states.

This formulation of overcoming conflict seems consilient with Freud's (1915/1963) claim about the conflict that underlies repression:

> *"Let us confine ourselves to the clinical experience we meet with in the practice of psychoanalysis. We then see that the satisfaction of an instinct under repression is quite possible; further, that in every instance such a satisfaction is pleasurable in itself, but is irreconcilable with other claims and purposes; it therefore causes pleasure in one part of the mind and 'pain' in another. We see then that it is a condition of repression that the element of avoiding 'pain' shall have acquired more strength than the pleasure of gratification (p. 105)."*

Stated in the language of the current paper, the pleasure of gratification of an instinct is here being understood as the "pleasure" of reducing the FE of a subsystem (a "part of the mind" as Freud suggests, in the current example, one of the prototype emotion systems). However, despite this important consilience with Freud's perception of repression, the mechanism of defense described here also has some important differences from the accepted description of conflict and repression in Freud's work. Specifically, what is missing from this description of the dynamic unconscious is almost the entire description from Freud's (1923/1961) structural theory, which is the role of a repressive action from the ego to avoid anxiety. This might lead one to suggest that the unconsciousness described above (due to not activating a dominant prediction) is not the same thing as that described by Freud. This is correct, it is not the same. The formulation of conflict and unconsciousness being presented here demand a different metapsychological assumption from that articulated in Freud's work.

Specifically, what is needed to incorporate a FEP-inspired description of conflict, repression and the unconscious within Freudian metapsychology is a systems-based epistemology which suggests that all the key mental processes of interest to psychoanalysis must be located within a hierarchy of organization, and must themselves be founded upon and constrained by processes at lower levels of the hierarchy (Connolly and van Deventer, 2017). This was expressed best in Grobbelaar (1989) as follows:

> *"As it stands now, Freud's formulation of the process of censorship defines it as an ad hoc defensive maneuver by one system, the ego, against another system, the unconscious, to stop dangerous elements (dangerous to the organization of the ego) from entering the ego. One should rather formulate from the bottom to the top, that is, in a theoretical sense. One should begin by defining the inherent qualities in the lower-order elements which … make it impossible for them to be taken up in a higher order system … (p. 142)."*

He elaborated on this further:

> *" … the principles determining the perception of thoughts will be inherent in the thoughts themselves. Stated differently, if the organization of the ideational domain does not allow for the representation of certain ideas, thoughts or memories, then they cannot become conscious (p. 139–140)."*

The incredible value that the FEP (and more specifically the current formulation of expected free energy of competing policies of action) can have for psychoanalysis is that it offers precisely such an explanation that offers a hierarchical description of the processes and also implies that defense (or repression) as a process must have its origin in process at a lower level of organization than consciousness, as described next.

To view precisions in terms of a hierarchical arrangement of functions in the brain, the precisions associated with a particular level of functional hierarchy are essentially determined by activity and structure at higher levels of hierarchy than that level at which the conflict takes place. In one regard, this refers to the range of possible states that can be encoded by activity at the higher levels. In terms of complexity of organization, it means that lower levels of complexity of encoding at the superordinate level result in higher levels of precision associated with surprise from subordinate levels (Mathys, personal communication, 14 July 2017). With regard to resolving conflict, this means that the generative model at a higher order comes to encode a more limited set of possible states with regard to one of the conflicting neural systems.

This is also important with regard to conscious experience. If we suggest as Hobson et al. (2014) have, that consciousness might refer to the process of inference at higher levels of brain hierarchies, then this might imply that the person in our example would likely only consciously experience the fear-related response, and be less aware of anger in their response. Note that it is possible that "automated" motor responses to anger, which may be triggered at levels of processing far below consciousness, such as forming a fist or clenching teeth, may nonetheless persist in the person's behavior. However, we might expect that they are usually not attended to by the person, though it is these behaviors which may typically be pointed out to a client by a psychoanalytically-oriented psychotherapist.

In this hierarchical setting, it is possible that policies unfold at different hierarchical levels, where the precision of prior preferences—that underwrite expected free energy—are supplied by supraordinate levels. In what follows, I will use precision as a shorthand for the precision of various prior preferences that determine policy selection at each and every level of inference. This comfortably accommodates the above phenomenology. For example, I can select low level (automatic and autonomic) policies that entail "fist clenching" and yet ignore this evidence that I am "angry" at a higher level of inference, if there is a more plausible (or "important") explanation or policy at hand (e.g., "I must do this to avoid being frightened"). In short, the precision-afforded prior preferences throughout the hierarchy play a crucial role in contextualizing the evidence for my current narrative and course of action—that will necessarily entail competition and ambiguity at each and every level[2].

The present formulation offers a potential formalization of defense related to conflict. Here, defense must refer to constraint reflected in the encoding at a level of functional hierarchy superordinate to that of action selection described earlier that results in an imperative to favor one policy of action over others. Referring to the above example, this could mean favoring a policy of action driven by the "FEAR" system rather than that of "RAGE." The result is that in future situations similar to that which triggered the conflict, we might expect the child from the "strange situation" example to be more likely to show a fear-related response and cling to the mother, and less likely to show an angry response.

[2]For a technical illustration of this sort of deep hierarchical inference see Friston et al. (2017b) which describes simulations dealing with the simple act of reading; where choosing which page to look at contextualizes, and informs choosing which sentence to sample, which contextualizes and informs, choosing the word to fixate on—and so on).

THE ENTRENCHMENT AND PROGRESSIVE COMPLEXITY OF DEFENSE MECHANISMS THROUGH DEVELOPMENT

This formalization of defense has implications for the further development of the person. The development of a person is characterized as the emergence of successively higher levels of hierarchical organization in the brain, as well as the progressive increase of complexity within those levels.

Further, following Grobbelaar (1989), and Connolly and van Deventer (2017), we could state that the constraints that operate at one level of a hierarchy must be reflected at higher levels of recursion. This means that the more "rigidly" encoded precisions of the defense must act as a constraint to the further development of the organism, including through hierarchically superordinate levels. This is formally similar to Freud (1912/1963) statement in "The Dynamics of the Transference":

> *"Now our experience has shown that of these feelings which determine the capacity to love only a part has undergone full psychical development; this part is directed toward reality and can be made use of by the conscious personality, of which it forms a part. The other part of these libidinal impulses has been held up in development, withheld from the conscious personality and from reality, and … may remain completely buried in the unconscious so that the conscious personality is unaware of its existence (p. 106)."*

This developmental aspect of defense is a critical element of a psychoanalytic view of the person. It is a common assumption within a psychoanalytic approach that a wide range of diverse adult behaviors are nonetheless thematically related to one another as being underpinned by a common defensive operation which has its origin in a critical event or situation from early childhood (Greenson, 1967). Returning to our example, we might find that the child who formed an imperative toward action policies related to fear rather than anger in situations that activate both (and that imperative has constrained further development) may as an adult exhibit a general inhibition of angry responses, and privileging of fearful behaviors in situations that call for both. For example, in a future adult relationship, when the person's partner arrives hours late for a meeting with little explanation or empathy, the person may appear to excessively seek reassurance rather than (consciously) expressing anger. Similarly, they may usually advise friends to behave in a placatory manner instead of an angry one when feeling mistreated by their partners. They may feel uncomfortable when observing someone expressing anger at their partner over perceived neglect, and try to avoid being exposed to situations where they might observe this behavior. Though these behaviors occur in different settings and situations, and reflect a complexity of influences, the present formulation attempts to show that they may indeed be traceable to a constraint on the developing hierarchical structure of the generative model that emerged at an early age, and became foundational to an emerging structure of perceptual (and action) prediction.

This progressive development of complexity of behavior and psychic activity associated with the defensive encoding of precisions is connected with clinical theory in psychoanalysis where it is proposed that the unconscious material comes into association with other elements of structure in the psyche, resulting in a diversification of expression of the related defense. Freud (1915/1963) describes this process from a clinical perspective:

> "… repression proper [emphasis in original], *concerns mental derivatives of the repressed instinct-presentation, or such trains of thought as, originating elsewhere, have come into associative connection with it. …We have to consider… the attraction exercised by what was originally repressed upon everything with which it can establish a connection. Probably the tendency to repression would fail of its purpose if these forces did not cooperate, if there were not something previously repressed ready to assimilate that which is rejected from consciousness. …we are inclined to … forget too readily that repression does not hinder the instinctual presentation from continuing to exist in the unconscious and from organizing itself further, putting forth derivatives and instituting connections (p. 106)."*

The above quote focuses on the progressive association of what is repressed with other elements of the psyche rather than the constraint related to the defense. In the current formulation, we might focus on the other side of the coin which is the increased precision of the now-dominant action policy, which must come to be applied to all new situations which trigger the previously conflicting state, such that the predictions associated with the dominant defensive response become ever more elaborated, through the ordinary development of the individual.

INERTIA AND THERAPEUTIC RESISTANCE

From the perspective of active inference, the tendency to increase the complexity of generative modeling within a creature's comfort zone can be understood in terms of free energy minimization: in the same way that expected free energy can be divided into epistemic and pragmatic parts, the free energy itself can be expressed as accuracy minus complexity (Hopkins, 2016). This means that as the generative model is optimized (i.e., learned through experience), it will try to provide more and more accurate explanations for its sensations. This will necessarily incur a complexity cost. Provided the accuracy increases—with learning—to a greater extent than the complexity, free energy will continue to decrease. This accuracy of the model is also dependent in part on the relative plasticity of the environment, such that the person can shape the environment in such a way that the generative model is accurate. This means that if a creature can find and construct its own "econiche" (an environment that fits and sustains the predictions of their model of the world), that generative model will increase its complexity only up until a point that there is no further gain (in terms of accuracy). Beyond this point, the phenomenon of (statistical) overfitting emerges. This corresponds to a failure to generalize the model to slight changes in the data, which means our model is no longer optimal to explain the normal levels of variation of data in our econiche. In this case, the generative model then appears to resist further change, provided the environment adequately supports the model as it is.

This process is important to psychoanalysis as it could explain the tendency to maintain a particular defensive constraint in the encoding of precisions of prior preferences that shape expected free energy and ensuing policies. These preferred outcomes specify states that become attractor states; namely; states to which the system is attracted; thereby maintaining its own organization and remaining within particular boundary parameters. The notion of self-maintenance and attractor states speaks directly to the premise of the free energy formulation—in the sense that the raison d'être for minimizing free energy is to establish and maintain experienced states within some attracting set; specified largely by prior beliefs (Friston, 2013). This theme emerges at many levels in self-organization; ranging from self-assembly in computational chemistry and molecular biology (Cademartiri et al., 2012; Friston et al., 2015a), through to autopoiesis (self-creation) in biological self-organization (Maturana and Varela, 1980; Thompson and Varela, 2001). The key point here is that if the priors that anchor the choice of action policies (to resolve conflict) become too entrenched, a particular, self-fulfilling, self-sustaining pattern of behavior emerges. Indeed, if this pattern involves placatory or reassuring behavior in the face of apparently devaluing behavior from others, one can imagine a particular personality phenotype (or ego-structure) that avoids aggressive behaviors within relationships altogether, to the extent that this behavior is successful and sustainable in meeting the expectations of the generative model.

The argument here is that "inertia" may reflect an entrenchment of prior beliefs that are sculpted by the imperative to avoid conflict and, in epistemic terms, the implicit uncertainty. Again, we see the imperative of reducing expected free energy or uncertainty in driving both behavior and the prior beliefs that underwrite that behavior. What this means regarding the present formulation is that as the generative model of the person continues to develop in complexity and hierarchical organization, so the constraint of precisions regarding action policies related to conflict come to be proportionally reflected in the generative model as well (though noting that it is also depending on the plasticity of the environmental niche as well). This means that the free energy cost of altering the precision of preferences that underwrite policy selection also increases with development.

This is important to understanding the tendency toward resistance in the therapeutic situation as well. Essentially, the task of the psychoanalytic therapist is to help the client reduce the relative precision of the dominant response, and allowing an increased precision of the opposing response such that it can activate conscious-level inference, and thereby have greater flexibility in behavior. However, this is tantamount to a kind of "attack" on the attractor state of the generative model. The inertia of the present encoding of precisions as described above clarifies the intensity with which the client avoids this conflicting information in terms of actions taken by the person to prevent the progress of the therapeutic activity (Nord et al., 2017 have recently connected the vigor of avoidance activities with the predictions regarding the likelihood of catastrophe, providing interesting possibilities for a predictive coding-informed perspective on avoidance, and potentially therefore, resistance).

However, this resistance or inertia against psychological change is not only evident in the actions the person takes to prevent the therapeutic progress, but also the updating of the generative model in such a way that the new information is "explained away" in an intellectual sense. This may be linked to the common observation in therapy where the therapist's interpretation, rather than facilitating the client toward meaningful restructuring of their ego defenses, rather just becomes another link in the chain of the client's defenses. The client understands or may even agree with the therapist, but meaningful change does not take place. The client is able to generate new verbalizations and thought in response to the therapist's efforts that merely support the generative model rather than driving change. The present formulation describing the progressive increases in complexity of the "defensive" generative model helps make sense of this phenomenon as well.

SOME IMPLICATIONS FOR PSYCHOANALYTIC THEORY AND THERAPY

A useful element of the formulation presented in this paper is that it appears to address the problems related to the "signal" theory of unpleasurable discharge that Freud developed in "The Interpretation of Dreams" 1990/1991. Here, Freud was pressed to explain how the psychic apparatus could prevent the mind from thinking of or remembering psychic material that caused unpleasurable discharge without experiencing it first. He suggested that there is a preliminary release of unpleasure associated with psychic activity that acts as a signal to the preconscious gate that discharge will cause unpleasure. As suggested in Grobbelaar (1989) and Connolly and van Deventer (2017), this process was never founded upon a suitable explanatory framework. However, the present formulation using expected free energy accomplishes this task. In essence, the updating of the generative model after the first experience of conflict means that the conflict state itself becomes reflected at a superordinate level of organization through the altered precisions. The sensory stimuli which would previously have generated the conflict state of uncertainty now generates the defense state that privileges one response over another. An example of such a response might be an inhibitory response of the prefrontal cortex toward the limbic system, which now occurs without necessarily reexperiencing the initial conflict state, but is rather the result of a downward prediction encoded at a cortical level. In essence the conflict is now "predicted" and "resolved" through one stroke, through the precision weightings toward one pole of the conflict now avoiding the uncertainty of the conflict state. Certainly, the organism also learns to avoid stimuli that activate that state or the surprise related to it. As stated earlier, this may be a reason why psychoanalytic conflict is difficult to measure in imaging of adult brains due to the fact that the established inhibitory, repressive behaviors of the brain, may often succeed in preventing the full experience of conflict as it has been defined in this paper.

A last implication that will be examined here relates to the role of therapy in restructuring the generative model. In one sense, the therapist could just point out to the client that there are actions that they are motivated to perform, though they aren't aware of it. Freud addressed such a situation in "Wild psycho-analysis" (1910) where he suggested that simply telling the client that they have unconscious motives are likely to make the client uncomfortable as it activates the conflict around it. He felt that such direct statements without regard to the therapeutic process brought psychoanalysis into disrepute as clients made so uncomfortable by comments such as this were often vocal in their condemnation of professionals who made such statements toward them, though Freud also felt that in the long run they might ultimately be helped by such statements as they drew the client's attention to the difficulty, at least. However, in that same paper, he suggested that a more therapeutically effective response (that also protected the dignity of the discipline) took into account two factors. Firstly, the readiness of the client, in the sense that they themselves were "in the neighborhood of" recognizing the repressed motivations themselves (which implies a lower FE cost in terms of perceiving it), but also that the relationship between therapist and client had reached a certain stage of emotional closeness in their relationship. This last is critical in the sense that the intensity or nature of the relationship with the therapist somehow alters the computation made by the person in terms of precisions related to expected free energy. An important question for future work is to state exactly *how* the relationship achieves this change that allows a recalculation of precisions associated with the conflict situation and the prototype emotions often associated with these.

One idea worth considering is the psychoanalytic notion of containment as articulated by Bion (1963). Based on Klein's (1946) concept, containment refers to the idea that the painful emotions and anxieties experienced by a person can in a sense be reduced in a relationship with another person, through a projection of the painful experience into the other who is experienced as becoming (through projective identification) the "bad" parts of the self. This seems to reduce the intensity of the emotions activated, and make the feelings seem more manageable. An example would be a person managing feelings of anxiety at separating from a loved one by (wrongly) perceiving a loved one as being very anxious about them instead, and feeling contempt for the other's perceived dependency. In this way the other person "contains" the feelings of anxiety. A precondition of this projective identification is the experience of the other as "good" in the sense that they can tolerate the negative emotions and be expected not to retaliate or abandon the person—in this sense the relationship is perceived as safe, despite these projective identifications of "bad" emotions. While the concept of containment from a FEP perspective requires a detailed treatment of its own, we could suggest that this perceived safety of the relationship must surely alter the perceived consequences of acting on emotions that might otherwise be repressed. Here we use again the example from Hopkins (2016) of the child who showed a pattern of fear responses (e.g., seeking reassurance) in key relationships while angry responses appeared absent, and developed into an adult who repressed angry responses in primary relationships. While the fear of expressing anger may have overwhelmed the young child, the adult in therapy who could feel anger toward a perceived abandonment by the "safe" therapist can learn to anticipate a far lower free energy cost of acting on that anger toward the therapist. This also forms the basis of Freud's (1912/1963) understanding of the therapeutic mechanism of transference, where the repressed emotion can be felt toward the therapist, allowing for it to achieve consciousness where it might otherwise not have. However, these remarks require more rigorous development in future than given here.

CONCLUSION

The present paper has examined the Freudian notion of conflict, and assumed that this part of the theory requires a quantitative explanatory framework. After highlighting the failed explanation of Freud's energic theory, a formulation around expected free energy was shown to be a viable alternative to the energic theory. This formulation proposes a computational or mathematical formalization of conflict, which refers to the situation of relatively equivalent expected free energy of a number of actions under competing polices. This formulation also offers a formalization of defense as a recalibration of precisions at a hierarchically superordinate level of organization. This defensive organization is viewed as constraining the further development of the generative model, such that it maintains an attractor state characterized by the defensive operation, though it manifests in behavior in a complex and multi-faceted way. Implications of this formulation were explored, with the ongoing question of the role of the therapeutic relationship identified as an ongoing question.

The free energy principle and predictive coding presents an exciting opportunity to psychoanalysis, in that core conceptual foundations of psychoanalysis can be re-examined in the light of predictive coding, not only in order to demonstrate the viability of the basic theory of psychoanalysis relative to a foundation in systems theory and neuroscience, but also to consider how the theory may need to be recast in a newer systems-based language that makes these links. Although some way off at this stage, one of the practical utilities of having a formal theory is that one can simulate active inference and dyadic interactions. In principle, this makes it possible to create *in silico* psychotherapy and provide proof of principle of some of the dynamics that one might hypothesize. Such work may also eventually influence the clinical practice and training of psychoanalytic theory.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Ainsworth, M. D. S., and Bell, S. M. (1970). Attachment, exploration and separation: illustrated by the behaviour of one-year-olds in a strange situation. *Child Dev.* 41, 49–67. doi: 10.2307/1127388

Anderson, M. C., and Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends Cogn. Sci.* 18, 279–292. doi: 10.1016/j.tics.2014.03.002

Anderson, M. C., Ochsner, K. N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., et al. (2004). Neural systems underlying the supression of unwanted memories. *Science* 303, 232–235. doi: 10.1126/science.1089504

Basch, M. F. (1976). Psychoanalysis and communication science. *Annu. Psychoanal.* 4, 385–421

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Berlin, H. A., and Montgomery, J. (2017). "Neural basis of intraspsychic and unconscious conflict and repetition compulsion," in *Psychoanalytic Perspectives on Conflict,* eds C. Christian, M. N. Eagle, and D. L. Wolitzky (New York, NY: Routledge), 260–278.

Bion, W. R. (1963). *Elements of Psycho-Analysis.* London: William Heinemann.

Bowlby, J. (1969). *Attachment and Loss: Attachment*, Vol 1. New York, NY: Basic Books.

Bowlby, J. (1973). *Attachment and loss: Separation, anxiety and anger*, Vol 2. New York, NY: Basic Books.

Breuer, J., and Freud, S. (2004). *Studies in Hysteria*. New York, NY: Penguin Books (Original work published in 1895).

Bronfman, E., Parsons, E., and Lyons-Ruth, K. (1999). *Atypical Maternal Behavior Instrument for Assessment and Classification (AMBIANCE): Manual for Coding Disrupted Affective Communication.* Harvard University Medical School, unpublished manual.

Cademartiri, L., Bishop, K. J., Snyder, P. W., and Ozin, G. A. (2012). Using shape for self-assembly. *Philos. Trans. Ser. A Math. Phys. Eng Sci.* 370, 2824–2847. doi: 10.1098/rsta.2011.0254

Carhart-Harris, R. L., and Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain* 133, 1265–1283. doi: 10.1093/brain/awq010

Cittern, D., Nolte, T., Friston, K., and Edalat, A. (2018). Intrinsic and extrinsic motivators of attachment under active inference. *PLoS ONE* 13:e0193955. doi: 10.1371/journal.pone.0193955

Colibazzi, T. (2016) July. "Dealing with conflict: too little or too much? perspectives from the psychosis prodrome," in *Paper Presentation at the 17th Annual Neuropsychoanalysis Congress* (Chicago, IL).

Connolly, J. P. (2016). *Principles of Organization of Psychic Energy Within Psychoanalysis: A Systems Theory Perspective*. Unpublished doctoral thesis, University of South Africa, Pretoria.

Connolly, P. (2017) July. "Expected free energy formalizes conflict and defense," in *Poster Session Presented at 18th Annual Congress of the Neuropsychoanalytic Society* (London).

Connolly, P., and van Deventer, V. (2017). Hierarchical recursive organization and the free energy principle: from biological self-organization to the psychoanalytic mind. *Front. Psychol.* 8:1695 doi: 10.3389/fpsyg.2017.01695

Dehaene, S., Artiges, E., Naccache, L., Martelli, C., Viard, A., Schürhoff, F., et al. (2003). Conscious and subliminal conflicts in normal subjects and patients with schizophrenia: the role of the anterior cingulate. *Proc. Nat. Acad. Sci. U.S.A.* 100, 13722–13727. doi: 10.1073/pnas.2235214100

Dutton, D. G., and Aron, A. P. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. *J. Pers. Soc. Psychol.* 30, 510–517. doi: 10.1037/h0037031

Freud, S. (1911/1963). "Formulations regarding the two principles of mental functioning," in *The Collected Papers of Sigmund Freud Volume 6*: *General Psychological Theory: Papers on Metapsychology*, ed P. Rieff (New York, NY: Collier Books), 21–28.

Freud, S. (1912/1963). "The dynamics of the transference," in *The Collected Papers of Sigmund Freud: Therapy and Technique*, ed P. Rieff (New York, NY: Collier Books), 105–115.

Freud, S. (1915/1963). "Repression," in *The Collected Papers of Sigmund Freud: General Psychological Theory: Papers on Metapsychology*, Vol. 6, ed P. Rieff, (New York, NY: Collier Books), 104–115.

Freud, S. (1920/1955). "Beyond the pleasure principle," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XVIII (1920-1922): Beyond the Pleasure Principle, Group Psychology and Other Works* ed J. Strachey (London: Hogarth Press), 1–64.

Freud, S. (1923/1961). "The Ego and the Id," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XIX (1923-1925): The Ego and the Id and Other Works,* ed J. Strachey (London: Hogarth), 3–66.

Freud, S. (1950). "The project for a scientific psychology," in *Pre-Psycho-Analytic Publications and Unpublished Drafts The Standard Edition of the Complete Psychological Works of Sigmund Freud* (1886-1899), ed J. Strachey (London: Hogarth Press), 281–399.

Freud, S. (1990/1991). *The Interpretation of Dreams*. Harmondsworth: Penguin.

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475. doi: 10.1098/rsif.2013.0475

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017a). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., Levin, M., Sengupta, B., and Pezzulo, G. (2015a). Knowing one's place: a free-energy approach to pattern regulation. *J. R. Soc. Interface* 12, 1–12. doi: 10.1098/rsif.2014.1383

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015b). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi: 10.1080/17588928.2015.1020053

Friston, K. (2009). The free energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. J. (2010). A free energy principle for the brain. *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle of the brain. *J. Physiol.* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017b). Deep temporal models and active inference. *Neurosci. Biobehav. Rev.* 77, 388–402. doi: 10.1016/j.neubiorev.2017.04.009

Gray, J. R., Bargh, J. A., and Morsella, E. (2013). Neural correlates of the essence of conscious conflict: fMRI of sustaining incompatible intentions. *Exp. Brain Res.* 229, 453–465. doi: 10.1007/s00221-013-3566-5

Greenson, R. R. (1967). *The Practice and Technique of Psychoanalysis.* New York, NY: InternationalUniversities Press.

Grobbelaar, P. W. (1989). *Freud and Systems Theory: An Exploratory Statement.* Unpublished doctoral dissertation, Rand Afrikaans University, Johannesburg.

Hartmann, H. H. (1964). *Essays on Ego Psychology.* New York, NY: International Universities Press.

Hobson, J. A., Hong, C. C., and Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Front. Psychol. Cogn. Sci.* 5:1133. doi: 10.3389/fpsyg.2014.01133

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Holt, R. R. (1962). A critical examination of Freud's concept of bound vs. free cathexis. *J. Am. Psychoanal. Assoc.* 10, 475–525.

Hopkins, J. (2012). "Psychoanalysis, representation and neuroscience: the Freudian unconscious and the Bayesian brain," in *From the Couch to the Lab: Psychoanalysis, Neuroscience and Cognitive Psychology in Dialogue*, eds A. Fotopoulu, D. Pfaff, and M. Conway (Oxford: Oxford University Press), 230–265.

Hopkins, J. (2015). "The significance of consilience: psychoanalysis, attachment, neuroscience and evolution," in *Psychoanalysis and Philosophy of Mind: Unconscious Mentality in the 21st Century,* eds S. Boag, L. A. W. Brakel, and V. Talvitie (London: Karnac), 47–137.

Hopkins, J. (2016). Free energy and virtual reality in neuroscience and neuropsychoanalysis: a complexity theory of dreaming and mental disorder. *Front. Psychol. Cogn. Sci.* 7:922. doi: 10.3389/fpsyg.2016.00922

Horowitz, M. H. (1977). The quantitative line of approach in psychoanalysis: a clinical assessment of its current status. *J. Am. Psychoanal. Assoc.* 25, 559–579. doi: 10.1177/000306517702500303

Howe, G. (2011). *Attachment Across the Life Course.* New York, NY: Palgrave MacMillan.

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Res.* 49, 1295–1306. doi: 10.1016/j.visres.2008.09.007

Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am. J. Psychiatry* 160, 13–23. doi: 10.1176/appi.ajp.160.1.13

Kessler, H., Schmidt, A. C., Hildenbrand, O., Scharf, D., Kehyayan, A., and Axmacher, N. (2017). Investigating behavioral and psychophysiological reactions to conflict-related and individualized stimuli as potential correlates of repression. *Front. Psychol.* 8:1511. doi: 10.3389/fpsyg.2017.01511

Klein, M. (1946). Notes on some schizoid mechanisms. *Int. J. Psycho-Anal.* 27, 99–110.

Krystal, J. H., Bremner, J. D., Southwick, S. M., and Charney, D. S. (1998). "The emerging neurobiology of dissociation: implications for treatment of posttraumatic stress disorder," in. *Trauma, Memory, and Dissociation,* eds J. D. Bremner and C. R. Marmar (Washington, DC: American Psychiatric Press), 321–363.

Levin, R., and Nielsen, T. (2009). Nightmares, bad dreams, and emotion dysregulation. *Curr. Dir. Psychol. Sci.* 18, 84–88. doi: 10.1111/j.1467-8721.2009.01614.x

Maturana, H. R., and Varela, F. (1980). *Autopoiesis and Cognition*. London: D Reidel.

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Front. Comput. Neurosci.* 10:56. doi: 10.3389/fncom.2016.00056

Nord, C. M., Prabhu, G., Nolte, T., Fonagy, P., Dolan, R., and Moutoussis, M. (2017). Vigour in active avoidance. *Sci. Rep.* 7:60. doi: 10.1038/s41598-017-00127-6

Panksepp, J. (1998). *Affective Neuroscience*. Oxford: Oxford University Press.

Panksepp, J., Siviy, S., and Normansell, L. (1984). The psychobiology of play: theoretical and methodological perspectives. *Neurosci. Biobehav. Rev.* 8, 465–492. doi: 10.1016/0149-7634(84)90005-8

Peterfreund, E., and Schwartz, J. T. (1971). *Information, Systems, and Psychoanalysis: An Evolutionary Biological Approach to Psychoanalytic Theory*. Psychological Issues Monograph 25/26. New York, NY: International Universities Press.

Pribram, K., and Gill, M. M. (1976). *Freud's Project Reassessed*. New York, NY: Basic Books.

Pushkarskaya, H., Smithson, M., Joseph, J. E., Corbly, C., and Levy, I. (2015). Neural correlates of decision-making under ambiguity and conflict. *Front. Behav. Neurosci.* 9:325. doi: 10.3389/fnbeh.2015.00325

Rapaport, D. (1960). The structure of psychoanalytic theory. *Psychol. Iss.* 2, 1–158.

Safyer, M. P. (2013). *When Good Enough Mothering Is Not Good Enough: A Study of Mothers' Secure Base Scripts, Atypical and Disrupted Caregiving and the Transmission of Infant Attachment Quality*. Unpublished Ph.D. thesis, University of Michigan.

Schmeing, J.-B., Kehyayan, A., Kessler, H., Do Lam, A. T. A., Fell, J., Schmidt, A.-C., et al. (2013). Can the neural basis of repression be studied in the MRI scanner? New Insights from Two Free Association Paradigms. *PLoS ONE* 8:e62358. doi: 10.1371/journal.pone.0062358

Schmidt, A., Diwadkar, V. A., Smieskova, R., Harrisberger, F., Lang, U. E., McGuire, P., et al. (2015). Approaching a network connectivity-driven classification of the psychosis continuum: a selective review and suggestions for future research. *Front. Hum. Neurosci.* 8:1047. doi: 10.3389/fnhum.2014.01047

Shevrin, H., Bond, J. A., Brakel, L., Hertel, R. K., and Williams, W. J. (1996). *Conscious and Unconscious Processes: Psychodynamic, Cognitive and Neurophysiological Convergences*. New York, NY: Guilford Press.

Shevrin, H., Snodgrass, M., Brakel, L. A. W., Kushwaha, R., Kalaida, N., and Bazan, A. (2013). Subliminal unconscious conflict alpha power inhibits supraliminal conscious symptom experience. *Front. Hum. Neurosci.* 7:544. doi: 10.3389/fnhum.2013.00544

Solms, M., and Turnbull, O. (2003). *The Brain and the Inner World: An Introduction to the Neuroscience of the Subjective Experience*. New York, NY: Other Press.

Swanson, D. R. (1977). A critique of psychic energy as an explanatory concept. *J. Am. Psychoanal. Assoc.* 25, 603–633. doi: 10.1177/000306517702500306

Thompson, E., and Varela, F. (2001). Radical embodiment: neural dynamics and consciousness. *Trends Cogn. Sci.* 5, 418–425. doi: 10.1016/S1364-6613(00)01750-2

van der Meer, M., Kurth-Nelson, Z., and Redish, A. D. (2012). Information processing in decision-making systems. *Neuroscientist*, 18, 342–359. doi: 10.1177/1073858411435128

White, G. L., Fishbein, S., and Rutsein, J. (1981). Passionate love and misattribution of arousal. *J. Pers. Soc. Psychol.* 41, 56–62. doi: 10.1037/0022-3514.41.1.56

Zepf, S. (2010). Libido and psychic energy – Freud's concepts reconsidered. *Int. Forum Psychoanal.* 19, 3–14. doi: 10.1080/08037060802450753

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 Psychology

CONCEPTUAL ANALYSIS
published: 27 September 2018
doi: 10.3389/fpsyg.2018.01674

Making Worlds in a Waking Dream: Where Bion Intersects Friston on the Shaping and Breaking of Psychic Reality

*Matthew John Mellor**

Kensington and Chelsea Psychotherapy Service, Central and North West London NHS Foundation Trust, London, United Kingdom

With the publication of Wilfred Bion's text 'Learning from Experience,' psychoanalysis was afforded a new schema for understanding the processes and implications involved in an infant's contact with their caregivers. As a result, our conception of some of the most fundamental phenomena of psychic life was significantly enriched. By proposing his theory of alpha-functioning, Bion mapped out how meaningful connexions to the internal and external worlds become established in the mind. In contrast, and through working clinically with psychotic patients, Bion revealed how these ties can catastrophically come undone. It is with these ideas, as well as their links to a corresponding set of neuroscientific constructs relating to the Markov blanket and principally developed by Karl Friston, that this paper is concerned. Through an investigation of the psychic functioning originally dubbed 'dream-work-alpha,' the paper's first section focuses on how Bion conceived of the creation of a 'contact-barrier' that allows for the differentiation of consciousness from an unconscious mind. Casting the ramifications of this organisation in sharp relief, the psychotic disorganisation of the contact-barrier is then explored. The discussion subsequently broadens to incorporate contemporary theories from free energy neuroscience that bear significant and illuminating relations to the psychoanalytic ideas espoused by Bion over half a century ago. Finally, through posing a series of three questions with accompanying discussions, a superimposition of these theoretical schemas is attempted. These suggestions directly address, (1) whether there is an intimate connexion between the interoceptive contact-barrier and the exteroceptive Markov blanket, (2) whether a disobjectalising of the contact-barrier may be reflected as a tear in the functional fabric of the Markov blanket, and (3) what the clinical implications are of working at the level of the projected surface. Ultimately, the aim of the paper is to expose relevant points of contact within and between the varying conceptual frameworks; frameworks that ultimately derive from disciplines that are both concerned with examining the underlying mechanisms of the mind-brain.

Keywords: Wilfred Bion, Karl Friston, Markov blanket, psychosis, alpha-function, contact barrier, free energy, dreaming

OPEN ACCESS

Edited by:
Jim Hopkins,
University College London,
United Kingdom

Reviewed by:
Jeremy Holmes,
University of Exeter, United Kingdom
Michel Botbol,
Université de Bretagne Occidentale,
France

***Correspondence:**
Matthew John Mellor
Matthew.mellor2@nhs.net

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *03 June 2018*
Accepted: *20 August 2018*
Published: *27 September 2018*

Citation:
Mellor MJ (2018) Making Worlds in a Waking Dream: Where Bion Intersects Friston on the Shaping and Breaking of Psychic Reality. Front. Psychol. 9:1674. doi: 10.3389/fpsyg.2018.01674

INTRODUCTION

In 1962, Wilfred Bion set out to deepen our conception of some of the most fundamental phenomena of psychic life by outlining the processes and implications involved in an infant's contact with their caregivers. Elaborating on Melanie Klein's notion of projective identification, Bion explored this contact in terms of the communication it facilitates (Segal, 2005). Going beyond parental contact and preverbal communication however, Bion's theorising of 'alpha-functioning' offered psychoanalysis a new schema with which to understand how meaningful connexions to the external and internal worlds become established in the mind. By contrast, the insights he gained through working clinically with psychotic patients reveal how these ties can catastrophically come undone. It is with these ideas, as well as their links to a corresponding set of neuroscientific constructs relating to the Markov blanket and principally developed by Karl Friston, that this paper is concerned.

Through an investigation of the psychic functioning originally dubbed "dream-work-alpha" (López-Corvo, 2003, p. 91), the focus of the paper's first section falls on how Bion conceived of the creation of a "contact-barrier" capable of "differentiating conscious from unconscious and maintaining the difference so established" (Bion, 1962a, p. 16). Casting the ramifications of this organisation in sharp relief, the psychotic disorganisation of the contact-barrier is then explored. The discussion subsequently broadens to incorporate contemporary theories from free energy neuroscience that bear significant and illuminating relations to the psychoanalytic ideas espoused by Bion over half a century ago. Finally, through posing a series of three questions with accompanying discussions, a superimposition of these theoretical schemas is attempted with a view to exposing relevant points of contact within and between the varying conceptual frameworks; frameworks that ultimately derive from disciplines that are concerned with examining the underlying mechanisms of "the same part of nature" (Solms, 2014).

AN OUTLINE OF ALPHA-FUNCTIONING

At the very heart of this synthesis is Bion's theory of alpha-functioning. For the infant faced with the task of developing a capacity for this, the prevailing external conditions play a pivotal role. Central to these conditions is the presence and temperament of the caregiver who, if able to effectively foster the infant's mind after birth, engages in a way of being described by Bion (1962b, p. 309) as "maternal reverie." In "good enough" (Winnicott, 1953, p. 94) conditions, this relationship allows the baby that possesses no "thought-thinking apparatus" to integrate their very first mental materials (Golse, 2003). Importantly for our concerns, Bion would come to classify these early emotional and sensory states as "beta-elements" that are liable for projection into the borrowed psyche of a "container" (Bion, 1962a, p. 6). Through the process of alpha-functioning, this containing figure is said to detoxify and transform beta-elements into "alpha-elements" that are capable of being assimilated by the infant (Golse, 2003). As Ferro (2011, p. 162) observes, the process equates to the conversion of proto-emotive chaos, into affectively meaningful representation.

Central to this exchange is the phenomenon of projective identification, the conventional definition of which is necessarily implicated provided the infant's primitive anxieties can be contained and transformed by the caregiver in the way described. By contrast to being seen simply as a "fantasy *in* the infant's mind" (Segal, 2005: my emphasis) where a psychological element is "displaced and relocated" (Laplanche and Pontalis, 1973, p. 349), the mechanism thus begins to resemble a search function that operates along the lines of a probe thrown into space (Pistiner De Cortiñas, 2011, p. 130). Where these projective probes encounter the kind of "transformational space" (Pistiner De Cortiñas, 2011, p. 130) that Bion had in mind, there exists the pregnant possibility of the baby's "proto-emotive chaos" being metabolised or "digested" (Bion, 1962a, p. 7). With this perspective, an active transferral is seen to take place both within and between the container and the contained. What's more, assuming this dyadic interaction can occur successfully and repeatedly, the infant ultimately stands to introject not only alpha-elements, but also their container's *very alpha-function* (Pistiner De Cortiñas, 2011, p. 130). Crucially, it is this process that's said to lead to the creation of a "contact-barrier": an internal "membrane" which proliferates where alpha-elements cohere, owes both its manifestation and structural integrity to the developmental trajectory that alpha-functioning inculcates, and which ultimately "marks the point of contact and separation between conscious and unconscious elements" (Bion, 1962a, pp. 17–22).

For Bion (1962a, p. 17) the capacity to transform the sense impressions related to an emotional experience, into alpha-elements is described as continuous in both sleeping *and* waking states. Indeed, the original name for alpha-functioning – "dream-work-alpha" – goes some way towards overtly acknowledging this fact. For the purposes of this paper, an ability to 'dream' (in inverted commas) will be invoked with Bion's meaning in mind; in other words that 'dreaming' reality works as a process of recording, assimilating and 'digesting' emotional experiences (Pistiner De Cortiñas, 2011, p. 136). Drawing on the theory that Freud had proposed in his seminal work on *The Interpretation of Dreams*, Bion suggests that the manifest content of a 'dream' should be considered as an enunciation that certain alpha elements are "constantly conjugated" (Pistiner De Cortiñas, 2011, p. 139). Understood as a context-sensitive adaption of Bion's (1962b, p. 306) "constant conjunction," a constant *conjugation* of alpha-elements reflects a clustered, connecting and combining agglomeration of the psychical products of dream-work-alpha. As such, one could say that we 'dream' the contact-barrier that creates "the distinction between the systems in the psychic apparatus" (Perelberg, 2005, p. 217); we *manifest* the "caesura"[1] that moves us out from being solely under the sway of a wishfulfilling pleasure principle.

[1] *Caesura*: Bion extended the term that he borrowed from Freud to a complex notion of gap, fissure, space, and bridge, having the function of both separating and communicating (Pistiner De Cortiñas, 2011, p. 129).

Turning briefly to address a further influence of Klein's on this schema, establishing a contact-barrier capable of allowing both separation and communication between the psychic systems was for Bion (1962a, pp. 23/24), closely connected with "the change from paranoid schizoid to depressive position and vice versa." As Bott Spillius et al. (2011, p. 78) note, it would take Bion's insights to develop the Kleinian concept of the psychic positions to the extent of it becoming common to think of "a moment-to-moment fluctuation between paranoid-schizoid and depressive states of mind." In positing the formula Ps↔D as reflecting the process of moving between psychic "disintegration and reintegration," Bion showed that "oscillations" between these positions are not only normal, but also required for the very "development of thoughts" (Bion, 1963, p. 35). As Britton (1998, p. 69) explains in a manner that resonates with the aforementioned description of why alpha-function is necessary, "thinking arises to deal with thoughts; thoughts require containing, naming and integrating." While 'D' involves producing a shape and containing it so as to imbue a meaning, the 'Ps' position must prevail for long enough for "the selected fact to emerge" (Britton, 1998, p. 69). Insofar as creative thinking resounds in the *coming into being* of integrating thoughts then, disintegration itself becomes an indispensable resource (Britton, 1998, p. 69).

Integrating these ideas into the present discussion and returning to *Learning from Experience*, we might suggest that the assimilation of a contact-barrier reflects a developmental "transition from a series of discrete particles or elements to a synthesis of these same elements" (Bion, 1962a, p. 24). In this sense, a transformation of beta-elements that had previously lacked "a capacity for linkage with each other" (Bion, 1962a, p. 22), into alpha-elements capable of constant conjugation, represents a binocular perspective on the transition from a world-view occupied by fragmented part-objects (Ps), to one where objects begin to be experienced ambivalently as separate and whole (D). It would fall to Segal (1957, p. 396) to highlight that this transition toward depressive integration has further consequences for an individual's ability to *use* symbols. As she writes, only when separateness is accepted in the working-through of the depressive position does the symbol become "a representation of the object rather than being equated with the object," the latter of which refers to the kind of symbolic equation synonymous with paranoid-schizoid functioning (Segal, 1981, p. 90). Recast in Bion's language, an ability to conjugate alpha-elements by 'dreaming' corresponds to this capacity to transform incoherent masses of stimuli and sensory impressions into symbolised "ideograms" or "pictograms" that may be used to register present and future experiences (Pistiner De Cortiñas, 2011, p. 136).

In order to arrive at a full understanding of dream-work-alpha, this concept of the ideogram must feature as an essential component. In short, it testifies to the fact that alpha-elements serve as both an input *and* output of the process thus described. As Bion (1962a, pp. 6/7) observes, whether asleep or awake, "emotional experiences have [. . .] to be worked upon by alpha-function before they can be used for dream thoughts [. . .]. If the patient cannot transform his emotional experience into alpha-elements, he cannot dream." In other words, an on-going capacity to transform latent dream thoughts into manifest dream content requires that assimilated nuclei of alpha-elements (ideograms) function as the unconscious imagos around which new conjugations of alpha-elements bind and cohere. While this process might be conceived of as latent dream thoughts connecting and cathecting at the contact-barrier, the overriding implication is that alpha-elements constitute both the raw material and yield of this "self-generating feedback system" (McGann, 1991, p. 15). Put yet more succinctly, alpha-functioning is therefore essentially autopoietic[2].

ALPHA-DYSFUNCTION AND PSYCHOTIC DISORGANISATION

In circumstances where individuals exhibit a profound lack of dream thoughts, a principal factor to be considered would be whether the early environment was able to provide "good enough" conditions. Without a containing presence helping to provide the infant with auxiliary support in their attempts to digest beta-elements, the individual may ultimately have difficulty entering into the self-perpetuating system of dream-work-alpha outlined above. As Bion (1962a, p. 6) describes them, beta-elements "are not amenable for use in dream thoughts but are suited for use in projective identification." In addition to being influential in producing acting-out, they are also used for the kind of thinking that depends on the "manipulation of what are felt to be things in themselves" (Bion, 1962a, p. 6). For the person incapable of transforming their experience symbolically, such a feeling of containing concrete things – rather than their images – can lead to the expectation of ideas behaving like sensory objects (De Masi, 2006, p. 20). As a result, these aspects of experience are liable to be split off and projected out of the mind in a manner that echoes the concreteness with which they're felt within.

While these characteristics and phenomena are consistent with Klein's account of the paranoid-schizoid position, a degree of elaboration is necessary if an adequately nuanced understanding is to be reached. Despite Klein's (1935, p. 145) assessment that infantile paranoid-schizoid anxiety is "comparable to the psychoses of adults," as Bion's work exposes, 'Ps' functioning can also be regarded as necessary for the healthy development of thoughts. It must therefore be emphasised that the Bionian notion of a disintegrative 'Ps' position that can engender future growth is exclusively applicable in instances where 'Ps' exists *in dialectical tension with 'D.'* In cases of profound early deprivation, a binary connexion between these positions may remain unestablished. Consequently, in the individual for whom the depressive position was never worked-through, there may be the pervasive lack of an ability to *use* disintegration resourcefully in the manner previously described; in other words, an incapacity to suspend attention with the kind of Keatsian "negative capability" that may creatively facilitate

[2]*Autopoiesis*: An autopoietic organism or machine is defined as one that "continuously regenerates and realizes the network of processes that produced them" (Maturana, 1973, p. 78).

the formation of new realisations and new states of mind (Keats, 1952, p. 383).

Important considerations are thus raised around the psychic mechanisms at play in situations where severe early deprivation *is* experienced and where Klein's comparison of paranoid schizoid anxiety to psychosis can be seen to resonate. As Winnicott hypothesises in *Fear of Breakdown*, when an infant endures extreme and traumatic conditions before they've developed sufficient perceptual apparatus to make sense of the overwhelming experience, psychotic defence organisations may be employed as a way to "short-circuit" the primitive agony (Ogden, 2014, p. 205). It's here that, for Winnicott, the seeds of psychosis are sown. Moreover, by not experiencing the breakdown in the "mother-infant tie" when it occurs, the individual creates a psychological state in which they live in fear of a breakdown that has already happened, but which was not experienced (Ogden, 2014, p. 205). For clarification, the "mother-infant tie" that Ogden refers to here is assumed to be the integral factor in the process of generating alpha-functioning that was explored in the opening section of this paper. In addition, the psychotic disavowal that perpetuates an individual's dissociation from the breakdown that was never experienced will, as Winnicott points out, be considered as a *communication* of the way the early environment failed (Winnicott, 1965, p. 128).

Should such a fundamental failure occur, the compulsion to avoid facing the pain of physical and emotional suffering can, Bion suggests, precipitate critically damaging consequences. As he puts it in his paper on the *Differentiation of the Psychotic from Non-psychotic Personalities,* the psychotic engages in the "minute fragmentation of the personality, particularly of the apparatus of awareness of reality" (Bion, 1957, p. 266). In other words, the very organs of emotional perception with which the experience would otherwise be registered may, in cases of extreme adversity, find themselves obliterated and eradicated. Moreover, such a wholesale attempt at neutralising pain can result in these fragments of the personality being expelled into external objects, where they become installed, often as a persecutory force (Bion, 1957, pp. 266–267). In Bion's (1957, pp. 268–270) terminology, such patients may consequently feel themselves "to be surrounded by bizarre objects" that carry a disturbingly "menacing presence."

By way of a clinical example of the bizarreness inherent in this psychic self-destruction, Bion, in his work *Cogitations,* recounts a patient who, "when unable to find the selected fact," externalises the terrifying experience through the enunciation "blood everywhere" (Pistiner De Cortiñas, 2011, p. 143). Bion's interpretive intervention in this instance was to convey that the patient had attacked their faculty for common sense which they thus saw spread everywhere as blood (Pistiner De Cortiñas, 2011, p. 143). What he was able to achieve with this insight and interpretation was the stemming of the tide by binding the spread fragments and formalising them into a scene (Pistiner De Cortiñas, 2011, p. 143). Cast in the language already prescribed, Bion lends his faculties and 'dreams' the "murder of common sense" (Pistiner De Cortiñas, 2011, p. 144) on behalf of his patient, thereby expressing his alpha-functioning and endowing the patient's experience with significance and meaning (Bell, 2011, p. 94).

When working with patients of a psychotic disposition, Bion emphasises the importance that the analyst is able to provide such auxiliary support by lending their faculties and 'dreaming' the session on behalf of the patient. In this respect – and contrary to the view held by both Freud and Immanual Kant for whom "the madman" was regarded as a "waking dreamer" (Stevens and Price, 1996, p. 229) – Bion saw the madman as *requiring* a waking dreamer in order to 'dream' the thoughts he can't. Furthermore, it is precisely this 'dreaming' (explored already as the capacity to consolidate alpha-elements), that provides the psychotic patient with invaluable containing tools for mental growth (Pistiner De Cortiñas, 2011, p. 140). In favourable circumstances, thoughts that previously lacked "a thinker" (De Masi, 2006, p. 51) may, within the carefully contained analytic situation, come to be circumscribed, symbolised and returned by the analyst to the agency from which they came.

Broadening the scope of the analytic technique under discussion and turning to a further Kleinian innovation, the process of playing, like 'dreaming,' can similarly be seen to produce constant conjugations of alpha-elements and facilitate the discovery of the selected fact (Pistiner De Cortiñas, 2011, p. 147). As Pistiner De Cortiñas (2011, p. 147) explains, playing within a clinical context can be used to harness feelings of guilt and criticism by encouraging patients to assign characters to the feelings previously experienced as "things in themselves." Through playful transformation, patients are thus presented with the possibility of opening imagined dialogues with their internal worlds. As will be shown, these principles have been demonstrated to have profoundly positive impacts in terms of helping individuals suffering with psychotic symptoms. Avatar therapy, a form of mental health treatment developed at University College London and Kings College London, constitutes a psychotherapeutic method that arguably both embraces and enhances precisely this approach. Moreover, having been explicitly designed for individuals experiencing auditory or visual hallucinations, both of which are "Scheiderian first-rank symptoms" (Tandon et al., 2013, p. 2) of schizophrenia, the therapy is particularly relevant to the concerns of this paper.

Originally invented by Julian Leff in 2008, avatar therapy entails patients working with a therapist to create virtual representations of their internal persecutors (Craig et al., 2018, p. 31). These avatars are constructed using specialised modelling software in order to bear as close a resemblance as possible to the visual and/or auditory characteristics of an individual's hallucinations (Craig et al., 2018, p. 33). Over a series of six weekly 50-min sessions, the patient engages in face-to-face work with the avatar, wherein the therapist facilitates a direct dialogue between the participant and avatar (Craig et al., 2018, p. 33). As the therapy progresses the avatar is engineered to gradually evolve from being persecutory to being supportive of the patient's strengths (Craig et al., 2018, p. 33). All sessions are recorded on an MP3 player that the patient takes away to use at home, particularly in instances where the voices are heard (Craig et al., 2018, p. 33).

As Craig et al. (2018, p. 31) observe, voice-hearers typically find themselves in a submissive role in relation to their voices,

a position that is characterised by feelings of inferiority and powerlessness which can reflect their experience of social relationships more generally. However, people who can establish a dialogue with their voices are often able to feel more power and control (Craig et al., 2018, pp. 31–32). A primary aim of avatar therapy is therefore to facilitate such a dialogue so that the voice-hearer can loosen the dominant (even omnipotent) grip of their voices (Craig et al., 2018, p. 31). In terms of the therapy's efficacy, a recent single-blind randomised controlled trial found that the treatment led to "a rapid and sustained reduction in the severity of auditory verbal hallucinations by end of therapy at week 12 that was significantly superior to that achieved by supportive counselling" (Craig et al., 2018, p. 38). It was in fact observed that multiple participants, many of whom had failed to respond to extended courses of antipsychotic medication, "reported a complete absence of voices during the preceding week at the week 12 assessment," with an even greater number experiencing such a cessation at 24 weeks (Craig et al., 2018, p. 37). Given that many of the participants in the study had been hearing voices for 20 years or more, such improvements should not be underestimated (Alderson-Day and Jones, 2018, p. 2).

Bringing these contemporary developments to bear on the Bionian theory explored, one could suggest that avatar therapy may function by allowing patients to 'dream' a persecutor that was previously only ever "coming out of the dark," to quote Samuel Beckett, one of Bion's own analysands (Oppenheim, 1994, p. 191). Furthermore (and in a manner close to how Pistiner De Cortiñas describes the clinical implications of 'playing'), the process would also seem to facilitate the transformation of persecutory feelings experienced as "things in themselves" through the secure opening of a dialogue. While the underlying mechanisms remain largely open to debate, there are grounds to consider that this 'playing' may expedite a kind of 'object creation' that promotes the formation of meaningful relations to an internal world previously felt as both illusive and intrusive. In other words, through imaginative simulation that's reinforced by the therapeutic apparatus, such hallucinatory presences may come to circumscribed and connected with, as opposed to being felt as uncontainably critical.

Thanks to the researchers conducting interviews with patients on the subject of their experiences while engaged in clinical trials of avatar therapy, there exists a substantial body of qualitative data from which to draw insights on the nature of voice hearing. One of the most significant findings insofar as this discussion is concerned is that, for many patients, the avatars and their voices come to represent feelings of low self-esteem that are related to past experiences of abuse and trauma (Craig et al., 2016, p. 49). Indeed for Romme et al. (2009, p. 25), in seventy per cent of individual cases the "voices are related to trauma and/or powerless making situations." This finding is further supported by recent large-scale, general population studies that have indicated that the relationship between childhood trauma, psychosis and schizophrenia is "a causal one with a dose effect" (Read et al., 2005). As such, there could be said to be renewed empirical validation for Winnicott's notion that psychotic symptoms should be regarded as a communication of the way in which the early environment failed. In less technical terms, for many of those experiencing the presence of voices that others don't perceive, that very presence may be imbued with a childhood trauma and its perpetrators.

Evidence concerning how psychosocial factors such as child abuse and neglect can affect an individual's likelihood of experiencing severe psychopathology raises further important considerations with respect to the nature and shape of the internal force that can, in such conditions, induce a profound fragmentation of the personality (Read et al., 2008, p. 235). When elaborating on the obliteration of psychic reality that's synonymous with the psychoses, the Kleinian superego represents a crucial piece to factor into the puzzle. In Klein's framework, the early superego is regarded as extremely severe, becoming less so in the process of development (Bott Spillius et al., 2011, p. 147). Crucially for this discussion, in pathological development, the severe early superego does not undergo modification; its pathogenic power may continue to be experienced in all its ruthlessness long beyond infancy. Bion would qualify Klein's thinking yet further with his notion of a primitive psychic agency that asserts itself with the ruthless effect of being *"opposed to, and destructive of, all links"* (Bion, 1959, p. 314). In Bion's understanding, this agency came to be explicitly defined as the "ego-destructive super-ego" and could be conceived of as responsible for attacking "links of emotion and reason between objects" (O'Shaughnessy, 2005).

Given the inherent difficulty in portraying the subjective experience of a fundamental disruption to meaning, locating material with which to apply this theory has a unique set of challenges associated with it. Nonetheless when consulting *The Centre Cannot Hold* – Elyn Saks's autobiographical novel that documents a life suffering with schizophrenia – one is offered a rare and compelling glimpse of the internal dynamics involved in a formidable superego unleashing a dissolutive wave. Speaking from the perspective of her 8-year-old self, Saks writes:

> *"My heart sinks at the tone of his [her father's] voice: I've disappointed him. And then something odd happens: My awareness (of myself, of him, of the room, of the physical reality around and beyond us) instantly grows fuzzy. . . I think I'm dissolving. . . like a sandcastle with all the sand sliding away in the receding surf. This is scary, please let it be over! Most people know what it's like to be seriously afraid. . . 'disorganisation' is a different matter altogether. . . One's centre gives way."* (Saks, 2007, pp. 12/13)

Having disappointed her father, an internalised version of whom comprises and configures her superego, Saks experiences the literal breaking down of physical and psychical reality. In theoretical terms, one might suggest that the passage illustrates the imposition and effect of a superego containing a "pure culture" (Freud, 1923, p. 53) of the death drive as described by André Green. More specifically, Saks's superego here triggers the release of a visceral force that "delinks, fragments, and unbinds" meaning (Reed and Baudry, 2005, p. 132). Furthermore, as has been argued to be at the crux of psychotic functioning, there would in this instance appear to be the terrifying "disorganisation" of the very apparatus with which meaning is conferred.

As Beckett (1957, p. 93) writes in *Endgame* when reflecting on the multitude of moments that make up an existence: "Grain upon grain, one by one, and one day, suddenly, there's a heap, a little heap, the impossible heap." For Saks, it's this "little heap" of a self that undergoes a collapse when the *"disobjectalizing function"* (Green, 2002, p. 646) of Green's re-envisioned death drive pervades the psyche with entropic force. Unlike the connecting, investing and objectialising creative power of Green's Eros that "links the infant with life, pleasure, the world of objects" (Reed, 2009, p. 5), in *The Centre Cannot Hold* we instead observe the "overtly psychotic" (Symington and Symington, 1996) disassembly of the psychic organs that ground the "centre" of the self. What's more, given what's already been explored in relation to the ways in which we come to bind psychic reality through the process of alpha-functioning, this disorganisation could be said to unravel and desolate the endopsychic contact-barrier between consciousness and the unconscious, thereby engendering "self-disappearance," "disinvolvement" and severing all comprehension and link with reality (Green, 2002, p. 646).

ON THE CONTACT-BARRIER WITH NEUROSCIENCE

As has been observed by many of those with mutual allegiances to psychoanalysis and the neurosciences, the Bionian views hitherto explored can be seen to fit comfortably with "those emerging from the new neuroscience" (Lipgar and Pines, 2003, p. 194). Going beyond the mere alignment of theory, however, Bion's insights into the nature of subjective experience arguably enhance the findings coming to light in this adjacent field (Lipgar and Pines, 2003, p. 194). Drawing on contemporary developments being pioneered at University College London – and chiefly utilising Karl Friston's investigations into the Markov blanket – much of the remainder of this paper will examine how these interfacing disciplines co-inform each other's understanding of the shaping and breaking of psychic reality. In advance of considering these synergies directly, an outline of the functional properties that Friston attributes to the Markov blanket is necessary in order to give body to the reflections.

For Friston (2013, p. 2), the Markov blanket has the fundamental property of inducing a "partition of states into internal and external states." Widely applied in probabilistic machine learning, it is this conceptual structure that, in a statistical sense, allows for boundaried distinctions to be drawn between different systems (Kirchhoff et al., 2018, p. 1). Central to our concerns here is the fact that the Markov blanket formalises the separation "between an inner and an outer environment" (Pezzulo and Levin, 2018, p. 32). In terms of a basic illustration, the cell represents an intuitive example of a living system with a Markov blanket; unless in possession of such a boundary, the cell would cease to exist as there would be no way of distinguishing it from the environment in which it lives (Kirchhoff et al., 2018, p. 2). Following this reasoning to its logical conclusion, evidence for any biological system is thus said to be contingent on it having a Markov blanket that facilitates the definition of inner from outer and which therein allows for the organism to be differentiated from that which it is not (Kirchhoff et al., 2018, p. 2). At the human level with which this paper is concerned, the Markov blanket is thus broadly conceived of as forming a sensory boundary, the activity beyond which consciousness has not been extended to.

Adding a further level of detail, internal states – the activity of which establishes the Markov blanket's existence – can themselves be subdivided into *sensory* and *active* states (Kirchhoff et al., 2018, p. 1). Sensory states are, in Friston's terms, defined as those that are caused by external states and which influence, but are not influenced by, internal states (Kirchhoff et al., 2018, p. 3). An example of this would be sensory information which is mediated by sensory states as it gets from the outside world, into the internal world (Friston, 2017). Active states, on the other hand, proceed in the opposite direction; they are caused by internal states and they influence, but are not themselves influenced by, external states (Kirchhoff et al., 2018, p. 3). An important consequence of this understanding is that the outer environment can only be seen "vicariously by the internal states, through the Markov blanket" (Friston, 2013, p. 2). It is for this reason that Friston refers to the Markov blanket as constituting a "veil" through which we infer the external causes of our sensory impressions (Friston, 2014a). Moreover, in addition to functioning as a metaphorical veil that discerns the sense impressions landing upon it, the Markov blanket also operates as a "projection screen" onto which are cast the habitual mechanisms (mediated by active states) that we use to *make sense of* the world (Friston, 2014a).

According to the exponents of free energy neuroscience, "active Bayesian inference" represents the chief mechanism by which we discern the sense impressions that come into contact with the Markov blanket (Friston, 2013, p. 1). Used for "calculating conditional probabilities," Bayesian inference involves creating and testing hypotheses, and updating beliefs in accordance with whether or not these predictions correspond to the data sampled (Joyce, 2008). Put simply, in instances of "prediction *error,*" what is expected to occur is invariably at odds with what is actually experienced. In light of such observations, a Bayesian system would necessarily revise itself in order to make more accurate predictions in the future. Cast in the terms already defined, this process corresponds to an organism attempting to preserve its existence by developing, maintaining and updating a "generative model" of its external environment (Kirchhoff et al., 2018, p. 5). Moreover, it is from within the periphery of the Markov blanket that the brain, functioning as a Bayesian machine, continually monitors the extent to which its internally constructed models accurately reflect the external reality that it stands in causal relation to. One of the most profound and over-arching implications of this with respect to human beings is that it revalidates Kant's notion that '*our manifest conscious image of ourselves as self-aware subjects of experience. . . is internal to our minds*' (Hopkins, 2012, p. 236).

Grounding all of this in psychoanalytic thinking, the way in which we come to know whether our generative model (and its constituent set of predictions about external states) is accurate or not is through affective feeling (Solms, 2014). Given that our perception of reality is all in the service of meeting our

needs in that reality, (which Freud wrote about in terms of it resulting in "an *experience of satisfaction*"), possessing an inaccurate generative model of the world would mean that needs remain unmet and affects come into play as a way of enforcing a revision to the model (Solms, 2014). When Friston therefore speaks about "minimising prediction error" and giving up on predictive models that don't correspond to external states, he's referring to Freud's reality principle, albeit in a different frame of reference (Solms, 2014). It is precisely this minimisation of prediction error – which results in a diminution of distressing affect – that ultimately sustains survival.

Adding another layer of detail so as to be able to comprehensively apply Bion's ideas, reducing prediction error (and therein conforming to the reality principle) are said to equate directly to the minimisation of "free energy" (Friston, 2014b). For Friston, minimising free energy is a defining trait of any biological system capable of preserving its existence over time (Friston, 2013, p. 2). Crucially and in the context of the preceding discussion, free energy is precisely the same quantity that is optimised (toward a minimum level) in Bayesian inference (Friston, 2013, p. 1). As such, an abundance of free energy – also known as "surprise" – would, in human beings, signal an individual making inaccurate predictions in relation to the world around them (Friston, 2014b). In psychoanalytic language, this translates as deficient reality-testing. In contrast, the process of resolving prediction errors and instigating effective reality-testing is, within Friston's paradigm, conceived of as involving the conversion of free energy into "bound energy" (Friston, 2014b). It is this "binding" that occurs within the boundary established by the Markov blanket and is described as fundamentally requiring the existence of "higher structures" in the organism (Friston, 2014b).

A principal consequence of this binding is that it allows the organism to operate in opposition to that which is "the long-term average of surprise": entropy (Friston, 2013, p. 2). By placing an "upper bound" on the entropy or dispersion of sensory states – (while simultaneously using those sensations to infer the external states of the world) – the organism in possession of a Markov blanket is thus able to "resist the second law of thermodynamics" (Friston, 2013, p. 2). Defined as the way in which isolated systems always evolve toward a state of maximum entropy, a resistance to this law within biological systems has the fundamental effect of allowing them to "preserve their functional and structural integrity" (Friston, 2013, p. 1). For systems that are incapable of resisting dispersion (and which therefore do not minimise free energy), the entropy of their sensory states "would increase indefinitely – by the fluctuation theorem," ultimately meaning that they "cannot exist" (Friston, 2013, p. 2). Moreover, not only does this vital ability to operate in opposition to entropy enable the continued existence of living systems, certain corollaries of it also facilitate their flourishing; as Friston (2013, p. 1) describes, evading dispersion allows for "homeostasis and a simple form of autopoiesis," the latter of which – as was alluded to earlier in this paper in relation to alpha-functioning – refers to a system capable of reproducing and maintaining itself. Appropriating the words of Zizek (2012, p. 467), we might therefore regard the Markov blanket as expediting "the gradual rise of order out of chaos."

SUPERIMPOSING THE CONCEPTUAL FRAMEWORKS

As was stated at the start of the paper, the concepts explored will now be examined from the perspective of where they intersect one another. Given the limitations of this paper however, compared to the scope of the material under discussion, one can only hope to present more questions than answers. As such, a series three suggestions will be posed with accompanying discussions, each of which will incorporate elements of understanding from the psychoanalytic and neuroscientific theories considered.

Is There an Intimate Connection Between the Interoceptive Contact-Barrier and the Exteroceptive Markov Blanket?

The suggestion here is that Bion's (1962a, p. 16) caesura which produces "ordered thought" by marking "the point of contact and separation between conscious and unconscious elements" (Bion, 1962a, p. 17), may potentially have a parallel correspondence with the Markov blanket that establishes "generalised synchrony" through using internal states to encode "events in the external world" (Friston, 2014a). Isolating the dynamics of the connexion more specifically, the infant's endopsychic contact-barrier which is constructed to facilitate the binding of beta-elements, could be conceived of as later projected out onto external reality, much in the way that Freud (1923, p. 26, my emphasis) describes the ego as the "*projection of a surface.*" Drawing on the ideas of Didier Anzieu, the suggested organisation might thus be said to resemble a "psychic envelope" (Jacobus, 2005, p. 9). In this frame, the contact-barrier that compounds interoceptive chaos by virtue of dream-work-alpha is projected out on external reality, forming a "visual dream-film" (Jacobus, 2005, p. 9) that functions mimetically to bind exteroceptive input. Through this view, we learn to work with the world having learnt to work with ourselves.

Such a conception of the "psychic envelope" is highly compatible with seeing the mind itself as a *container* (Hopkins, 2000, p. 8). Indeed, the English language has an abundance of analogies that instinctively pertain to precisely this understanding. From a forgetful person being described as having a "brain like a sieve," to an unstable individual being thought of as "out of their mind" or "having gone to pieces," there exists an entire family of metaphors that refer to a notion of the mind being circumscribed by containing boundaries (Hopkins, 2000, pp. 8–9). As has been demonstrated in relation to Bion's writing moreover, such individual psychic containment is made fundamentally possible by the dualism of a container-contained relation. While, we may therefore learn to work with the world having learnt to work with ourselves, as Vygotsky (1998, p. 170) observes, we actually *become ourselves through others*.

In terms of visualising the structural organisation of these interoceptive and exteroceptive boundaries, the infamous analogy of Plato's cave involves an imagined space and components that are particularly compatible with the proposed understanding (Solms, 2014). Without wishing to map the

analogy on in too concrete a manner, its philosophical implications could yet be seen to speak directly to some of Friston's basic proposals. In Plato's "strange image," multiple prisoners sit facing the wall of a cave, unable to move (Plato, 360 B.C.E). Behind them is situated a fire, in front of which is a walkway where unseen men patrol carrying statues (Plato, 360 B.C.E). For these hypothetical prisoners, all that's ever perceived are their own shadows and those cast by the statues "which the fire throws on the opposite wall of the cave" (Plato, 360 B.C.E). Despite the analogy progressing, it is this crucible that's of interest in this instance.

Of particular note is the extent to which the prisoners in Plato's theoretical cave are illustrative of the fact that, as Friston states, "we are only seeing our projections" (Solms, 2014). In fact, at the 2014 Sandler Conference, Friston went as far as to specifically describe the Markov blanket as a "projection screen" (Friston, 2014a). Quoting the 19th century physicist, philosopher and physician, Hermann von Helmholtz, he unambiguously stated that objects "are *imagined* in the field of vision to account for sensation" (Solms, 2014). Importantly for our concerns, Friston's model here echoes something fundamental to a psychoanalytic understanding of how individuals come to perceive the world around them: "they imagine a construction of the world," and yet this fantasy must account for actual sensation (Solms, 2014). Bringing in Bion, one might also add that these imagined objects are irrevocably coloured by phantasmagorias of the imagos that conjugate at level of the contact-barrier. In terms of a Platonic parallel, the statues that cast the shadows would be analogous to these imagos.

Might a Disobjectalising of the Contact-Barrier Be Reflected as a Tear in the Functional Fabric of the Markov Blanket?

As is evident from the significance of the processes associated with the Markov blanket, the structure carries a particular importance for the psychoanalytic understanding of how a person might experience a "loss of contact with reality," as Freud put it (Freud, 1924, p. 183). While Freud's terse description of psychosis remains an important gateway into considering the condition, bringing Bion together with Friston adds new dimensions to the understanding of the processes involved in such a 'loss of contact.' As Bion (1962a, p. 16) writes, "alpha-function, which makes dream possible. . . preserves the personality from what is virtually a psychotic state." Given the suggested interrelation between the structural layers of the "psychic envelope," it would therefore make sense to contend that damage to that which is the product of alpha-function – the contact-barrier – would correspondingly affect the operation of the projected surface: the Markov blanket.

As was explored in the context of the ego-destructive superego, damage to the contact-barrier may be the result of a Greenian death drive critically interrupting "relationships in the activity of the mind" (Green, 2010, p. 29). Furthermore, in circumstances where this psychic force "dissolves connections" (Reed, 2009, p. 5) at the level of the contact-barrier, then – *due to the projection of the surface* – we'd see the Markov blanket's composition and concordance necessarily implicated. As such, we can begin to perceive a direct relationship between a contact-barrier that is disobjectalised by a suffusion of the drive toward "neuronal inertia" (Freud, 1895, p. 296), and a Markov blanket that is 'torn' and thus ineffective.

In Friston's terms, a radical lack of the functionality that's otherwise induced by a structurally integral Markov blanket would result in a proliferation of unbound free energy. As discussed, this equates to massive prediction errors being made in relation to how the external world is expected to behave. Moreover, as Solms (2014) points out in no uncertain terms, "minimising prediction error *is* the reality principle." Therefore for the individual with an impaired capacity to minimise prediction error, "banging into" the aspects of reality that couldn't or wouldn't be sampled by the dysfunctional apparatus is likely to be a recurrent phenomenon (Solms, 2014). As Freud (1924, p. 185) puts the observation in his paper on *The Loss of Reality in Neurosis and Psychosis*, "in a psychosis the rejected piece of reality constantly forces itself upon the mind." As such, the psychotic experiences the return of the disavowed as the rejected pieces of reality repeatedly puncture the individual's distorted worldview (Quinodoz, 2005, p. 245). Of course, given what's been discussed, one could argue that our perception is always already distorted to some degree. Nonetheless, the rupturing of the contact-barrier through the "work of the negative" (Green, 1992, p. 586) would undoubtedly represent psychopathology of another calibre.

In extremely severe circumstances, an overwhelming profusion of free energy resulting from such a rupture would mean that the entropy of the individual's "sensory states would not be bounded" (Friston, 2013, p. 2). In other words, and by way of a more tangible example of how this might be experienced, Elyn Saks's aforementioned depiction of *"dissolving. . . like a sandcastle with all the sand sliding away"* gives a visceral impression of sensory states quite literally dissipating. Having been explored already in relation to Green's death drive, Saks's passage also speaks directly to Friston's assertion that an individual experiencing the wholesale disintegration of their Markov blanket would consequently fail in their attempts to minimise dispersion (Friston, 2014a). In such a scenario, an individual's "autopoietic maintenance" is crucially said to be at stake (Friston, 2014a). This process was similarly identified and explained in the first section of this paper as a key feature of alpha-functioning. Consequently, while correlation doesn't automatically mean causation, the numerous similarities between these interior and exterior surfaces would seem repeatedly to point to the existence of a reciprocal relationship between them.

What Are the Clinical Implications of Working at the Level of the Projected Surface?

Working at this level could be seen to be precisely what's afforded by the psychoanalytic technique of probing the patient's

transference. Utilised extensively in the clinical setting, the phenomenon is said to offer psychoanalysis "the inestimable service of making the patient's hidden and forgotten erotic impulses immediate and manifest" (Freud, 1912, p. 107). As Sandler (1976, p. 43) writes, the transference can therefore be regarded as comprising "a concealed repetition of earlier experiences and relationships" which are thus revived and projected onto the analyst. In terms of the question of which experiences and relationships we might expect to see revived and projected, Freud (1936, p. 18) identifies them as having "their source in early – indeed, the very earliest – object relations." Provided the analyst can contain the patient's projections through well-timed and accurate interpretations, however, there exists the possibility of these unconscious ways of relating being transformed into self-knowledge. In the words of Ferenczi (1933, p. 160), the patient that's contained in this way stands to "re-experience the past no longer as hallucinatory reproduction but as an objective memory."

Returning once more to avatar therapy, this pioneering method could be argued to represent a profound intensification of the transferential process that's facilitated by the classical analytic setting. By going beyond the transferral of internal objects, to the point of creating and projecting them into an avatar, this method of treatment facilitates a process whereby these hallucinatory and delusory presences are engaged in the form of externalised and newly recognisable imagos. Moreover, by contrast to the internal persecutory presence, an external avatar represents a persecutor securely contained and controlled. As was explored earlier in this paper, while these persecutors may be imbued with a childhood trauma and its perpetrators, in this specialised clinical scenario, they're disarmed of their ability to cause unrestricted damage. As such, Ferenczi's (1933, p. 160) psychoanalytic proposal that there must be a vital "*contrast between the present and the unbearable traumatogenic past*" would appear to have been here maintained.

Building on the suggested model of interrelated psychic surfaces, these moderated and virtually represented avatars – once introjected – could be said to function as clusters of contained affect around which new conjugations of alpha-elements may cohere. Viewed through the prism and language of Bion, the technique might thus be argued to allow for the formulation of more resilient coping strategies by enabling internal persecutors to be rendered increasingly accessible on the intra-psychic level of 'dreaming.' It's in this sense that capturing and engaging internal phenomena at the projected surface (in order to explore them securely and therapeutically), may have the corresponding effect of enhancing the person's interoceptive "sense organ for the apprehension of psychical qualities" (Freud, 1900, p. 574).

CONCLUSION

From Freud and Klein, to Green and Winnicott, Wilfred Bion's writings intersect and inform countless of the theories developed by his peers. Indeed, the contributions that Bion brought both to his own and other disciplines are far from static; they continue to unfold in new ways in accordance with emerging concepts and evolving modes of thought. As has been explored in this paper, the interface that exists between *Learning from Experience* and contemporary Fristonian neuroscience is the location of a particularly fruitful cross-fertilisation of ideas. While demonstrating the points where these rich veins of thought make contact has been the ultimate goal of this paper, due to their depth and complexity, there's undoubtedly more work to be done. Rather than drawing premature conclusions therefore, it is hoped that various openings have been indicated.

Reflecting more specifically, the esoteric ways in which mental life is shaped were considered as conspicuously revealed by an understanding of how psychic elements and functions move between container and contained. For Bion, it's as a result of this process that we form a receptive internal world in which conscious and unconscious elements are both separate and in communication. At the very crux of this organisation lies the ability to 'dream,' the implications of which centrally include the digestion of emotional experience and a propensity for regenerative growth. By contrast, failures of this capacity and the attempt to avoid the potentially painful perspectives it induces have been investigated in relation to severe psychopathology.

The continued relevance and flexibility of these theories testifies to the true extent of Bion's inter-disciplinary potential; nowhere is this clearer than when his concepts are brought into dialogue with Friston's proposals. As a result of this synthesis and through the applied use of the relationship between interoceptive and exteroceptive contact-barriers, profound therapeutic gains stand to be made. In this regard – and with technological advances making object creation at the level of the projected surface increasingly feasible – it's possible to conceive of further developments in how disturbed internal boundaries may be approached and reconstructed. As this work has endeavoured to show, it is through fostering these frontiers of consciousness that the dream of a contained unit self is made manifest.

AUTHOR CONTRIBUTIONS

MM is responsible for the work associated with the production of this paper.

FUNDING

Following a proportional fee waiver, the remainder of the publication fees associated with this paper have been funded by MM.

ACKNOWLEDGMENTS

This paper was first produced as a dissertation for an M.Sc. in Theoretical Psychoanalytic Studies at University College London (Mellor, 2014). The original work was supervised by Professor Peter Fonagy. It has since been revised for the purposes of inclusion in this issue of *Frontiers in Psychology*.

REFERENCES

Alderson-Day, B., and Jones, N. (2018). Understanding AVATAR therapy: who, or what, is changing? *Lancet Psychiatry* 5, 2–3. doi: 10.1016/S2215-0366(17)30471-6

Beckett, S. (1957). "Endgame," in *Samuel Beckett: The Complete Dramatic Works 2006* (London: Faber and Faber).

Bell, D. (2011). "Bion: the phenomenologist of loss," in *Bion Today*, ed. C. Mawson (Hove. Routledge), 81–101.

Bion, W. R. (1957). Differentiation of the psychotic from the non-psychotic personalities. *Int. J. Psycho-Anal.* 38, 266–275.

Bion, W. R. (1959). Attacks on linking. *Int. J. Psycho-Anal.* 40, 308–315.

Bion, W. R. (1962a). *Learning from Experience*. London: Tavistock.

Bion, W. R. (1962b). The psycho-analytic study of thinking. *Int. J. Psycho-Anal.* 43, 306–310.

Bion, W. R. (1963). *Elements of Psycho-Analysis*. London: Heinemann.

Bott Spillius, E., Milton, J., Garvey, P., Couve, C., and Steiner, D. (2011). *The New Dictionary of Kleinian Thought*. Hove: Routledge.

Britton, R. (1998). "Beyond the depressive position: Ps(n+1)," in *Kleinian Theory: A Contemporary Perspective 2001*, ed. C. Bronstein (London: Whurr), 63–76.

Craig, T., Rus-Calafell, M., Ward, T., Leff, J., Huckvale, M., Howarth, E., et al. (2018). AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. *Lancet Psychiatry* 5, 31–40. doi: 10.1016/S2215-0366(17)30427-3

Craig, T., Ward, T., and Rus-Calafell, M. (2016). "AVATAR therapy for refractory auditory hallucinations," in *Brief Interventions for Psychosis: A Clinical Compendium*, eds B. Pradhan, N. Pinninti, and S. Rathod (Cham: Springer), 41–54.

De Masi, F. (2006). *Vulnerability to Psychosis*. London: Karnac

Ferenczi, S. (1933). "Confusion of tongues between adults and the child," in *Final Contributions to the Problems and Methods of Psychoanalysis*, eds L. Aron and A. Harris (London: Hogarth Press), 156–167.

Ferro, A. (2011). "Clinical implications of bion's thought," in *Bion Today*, ed. C. Mawson (Hove: Routledge), 155–172.

Freud, A. (1936). *The Ego and the Mechanisms of Defence, 1979*. London: Hogarth Press.

Freud, S. (1895). "Project for a scientific psychology," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, eds J. Strachey, A. Freud, A. Strachey, and A. Tyson, Vol. 1. (London: Vintage), 281–391.

Freud, S. (1900). "The interpretation of dreams," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*: *The Interpretation of Dreams (First Part)*, eds J. Strachey, A. Freud, A. Strachey, and A. Tyson, Vol. 4 (London: Vintage).

Freud, S. (1912). "The dynamics of transference," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud: The Case of Schreber, Papers on Technique and Other Works*, eds J. Strachey, A. Freud, A. Strachey, and A. Tyson, Vol. 12 (London: Vintage), 97–108.

Freud, S. (1923). "The ego and the Id," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud: The Ego and the Id and Other Works*, eds J. Strachey, A. Freud, A. Strachey, and A. Tyson, Vol. 19 (London: Vintage), 1–66.

Freud, S. (1924). "The loss of reality in neurosis and psychosis," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*: *The Ego and the Id and Other Works*, eds J. Strachey, A. Freud, A. Strachey, and A. Tyson (London: Vintage), 181–188.

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface* 10, 1–12. doi: 10.1098/rsif.2013.0475

Friston, K. (2014a). "Consciousness and the bayesian brain," in *A Lecture Given at the Joseph Sandler Conference*, Frankfurt. Available at: https://www.youtube.com/watch?v=HeQfO4byFhg (accessed May 22, 2018).

Friston, K. (2014b). "Discussion of consciousness by surprise: the unconscious in psychoanalysis and cognitive science, by Mark Solms," in *A Lecture Given at the Joseph Sandler Conference*, Frankfurt. Available at: https://www.youtube.com/watch?v=B-CYWdpHROg (accessed June 03, 2018).

Friston, K. (2017). "Free energy principle," in *A Lecture Given to Serious Science*, ed. A. Babitsky (Moscow: Serious Science). Available at: https://www.youtube.com/watch?v=NIu_dJGyIQI (accessed July 30, 2018).

Golse, B. (2003). "Capacity for maternal reverie," in *The International Dictionary of Psychoanalysis*, ed. A. de Mijolla (Farmington Hills: The Gale Group, Inc.) Available at: https://www.encyclopedia.com/psychology/dictionaries-thesauruses-pictures-and-press-releases/maternal-reverie-capacity (accessed April 24, 2018).

Green, A. (1992). Cogitations. *Int. J. Psychoanal.* 73, 585–589.

Green, A. (2002). A dual conception of narcissism. *Psychoanal. Quart.* 71, 631–649. doi: 10.1002/j.2167-4086.2002.tb00020.x

Green, A. (2010). Sources and vicissitudes of being in D.W. Winnicott's Work'. *Psychoanal. Quart.* 79, 11–35. doi: 10.1002/j.2167-4086.2010.tb00438.x

Hopkins, J. (2000). "Psychoanalysis, metaphor, and the concept of mind" in *The Analytic Freud*, ed. M. P. Levine (London: Routledge).

Hopkins, J. (2012). "Psychoanalysis representation and neuroscience: the freudian unconscious and the bayesian brain" in *From the Couch to the Lab: Psychoanalysis, Neuroscience and Cognitive Psychology in Dialogue*, eds A. Fotopolou, D. Pfaff, and M. Conway, (Oxford: Oxford University Press).

Jacobus, M. (2005). *The Poetics of Psychoanalysis*. Oxford: Oxford University Press.

Joyce, J. (2008) "Bayes theorem," in *The Stanford Encyclopedia of Philosophy*, ed. N. Zalta (Stanford, CA: CSLI).

Keats, J. (1952). Negative capability and wise passiveness. *PMLA* 67, 383–390. doi: 10.2307/459816

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15:20170792. doi: 10.1098/rsif.2017.0792

Klein, M. (1935). A contribution to the psychogenesis of manic-depressive states. *Int. J. Psycho-Anal.* 16, 145–174.

Laplanche, J., and Pontalis, J. B. (1973). "The language of psycho-analysis" in *The International Psycho-Analytic Library,* Vol. 94. (London: The Hogarth Press and the Institute of Psycho-Analysis), 1–497.

Lipgar, R., and Pines, M. (2003). *Building on Bion: Origins and Context of Bion's Contributions to Theory and Practice*. London: Jessica Kingsley.

López-Corvo, R. (2003). "Dream-work," in *The Dictionary of the Work*, ed. W. R. Bion (London: Karnac).

Maturana, H. R. (1973). *Autopoiesis and Cognition: The Realization of the Living, 1980*. New York, NY: Springer.

McGann, J. (1991). *The Textual Condition*. Princeton, NJ: Princeton University Press.

Mellor, M. (2014). *Making Worlds in a Waking Dream: Where Bion Interects Others on the Shaping and Breaking of Psychic Reality*. Dissertation/Master's thesis, University College London, London. Available at: https://ucl.academia.edu/MatthewMellor

Ogden, T. H. (2014). Fear of breakdown and the unlived life. *Int. J. Psychoanal.* 95, 205–223. doi: 10.1111/1745-8315.12148

Oppenheim, L. (1994). *Directing Beckett.* Ann Arbour, MI: University of Michigan Press.

O'Shaughnessy, E. (2005). Whose bion? *Int. J. Psycho Anal.* 86, 1523–1528. doi: 10.1516/DDJD-MC5U-Y13N-YPUA

Perelberg, R. J. (2005). "Unconscious phantasy and après-coup: From the History of an Infantile Neurosis" in *Freud: A Modern Reader*, ed. R. J. Perelberg (London: Whurr), 206–223.

Pezzulo, G., and Levin, M. (2018). Embodying markov blankets. *Phys. Life Rev.* 24, 32–36. doi: 10.1016/j.plrev.2017.11.020

Pistiner De Cortiñas, L. (2011). "Science and fiction in the psychoanalytical field," in *Bion Today*, ed. C. Mawson, (Hove: Routledge), 121–152.

Plato. (360 B.C.E). "Book VII," in *The Republic*, trans. B. Jowett. Available at: http://classics.mit.edu/Plato/republic.8.vii.html

Quinodoz, J.-M. (2005). *Reading Freud: A Chronological Exploration of Freud's Writings*. Hove: Routledge.

Read, J., Fink, P., Rudegeair, T., Felitti, V., and Whitfield, C. (2008). Child maltreatment and psychosis: a return to a genuinely integrated bio-psycho-social model. *Clin. Schizophr. Relat. Psychos.* 2, 235–254. doi: 10.3371/CSRP.2.3.5

Read, J., Van Os, J., Morrisson, A P., and Ross, C A. (2005). Childhood trauma, psychosis and schizophrenia: a literature review with theoretical and clinical implications. *Acta Psychiatr. Scand.* 112, 330–350. doi: 10.1111/j.1600-0447.2005.00634.x

Reed, G. S. (2009). An empty mirror: reflections on nonrepresentation. *Psychoanal. Quart.* 78, 1–26. doi: 10.1002/j.2167-4086.2009.tb00384.x

Reed, G. S., and Baudry, F. D. (2005). Conflict, structure, and absence. *Psychoanal. Quart.* 74, 121–155. doi: 10.1002/j.2167-4086.2005.tb00203.x

Romme, M., Escher, S., Dillon, J., Corstens, D., and Morris, M. (2009). *Living with Voices: 50 Stories of Recovery*. Herefordshire: PCCS Books.

Saks, E. (2007). *The Centre Cannot Hold*. London: Virago.

Sandler, J. (1976). Countertransference and role-responsiveness. *Int. Rev. Psycho-Anal.* 3, 43–47.

Segal, H. (1957). "Symbolization" in *Kleinian Theory: A Contemporary Perspective*, ed. C. Bronstein (London: Whurr).

Segal, H. (ed.). (1981). "The function of dreams" in *The Work of Hanna Segal: A Kleinian Approach to Clinical Practice* (Lanham, MD: Jason Aronson).

Segal, H. (2005). "Alpha function," in *The International Dictionary of Psychoanalysis*, ed. A. de Mijolla (Farmington Hills, MI: The Gale Group Inc.).

Solms, M. (2014). "Discussion of consciousness and the bayesian brain, by Karl Friston," in *A Lecture Given at the Joseph Sandler Conference*, Frankfurt.

Stevens, A., and Price, J. (1996). *Evolutionary Psychiatry: A New Beginning*. London: Routledge.

Symington, J., and Symington, N. (1996). *The Clinical Thinking of Wilfred Bion*. London: Routledge.

Tandon, R., Gaebel, W., Barch, D. M., Bustillo, J., Gur, R. E., Heckers, S., et al. (2013). Definition and description of schizophrenia in the DSM-5. *Schizophr. Res.* 150, 3–10. doi: 10.1016/j.schres.2013.05.028

Vygotsky, L. S. (1998). "Cognition and language: a series in psycholinguistics," in *The Collected Works of L.S. Vygotsky: Child Psychology*, eds R. W. Rieber and A. S. Carton (New York, NY: Plenum Press).

Winnicott, D. W. (1953). Transitional objects and transitional phenomena: a study of the first not-me possession. *Int. J. Psycho-Anal.* 34, 89–97.

Winnicott, D. W. (1965). "The maturational processes and the facilitating environment," in *The International Psycho-Analytic Library*, ed. M. R. Khan, Vol. 64 (London: The Hogarth Press and the Institute of Psycho-Analysis), 1–276.

Zizek, S. (2012). *Less Than Nothing*. London: Verso.

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 Psychology

REVIEW
published: 11 October 2018
doi: 10.3389/fpsyg.2018.01861

The Epistemological Foundations of Freud's Energetics Model

Jessica Tran The[1,2,3]*, *Pierre Magistretti*[1,4,5] *and François Ansermet*[1,6]

[1] Agalma Foundation, Geneva, Switzerland, [2] Département of Psychoanalytical Studies, Paris 7 Diderot University, Paris, France, [3] Institute of Medical Humanities, Université de Lausanne, Lausanne, Switzerland, [4] Department of Psychiatry, Faculty of Medicine, Brain Mind Institute, Swiss Federal Institute of Technology in Lausanne, Lausanne, Switzerland, [5] Division of Biological and Environmental Science and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, [6] Department of Psychiatry, Faculty of Medicine, University of Geneva, Geneva, Switzerland

This article aims to clarify the epistemological foundations of the Freudian energetics model, starting with a historical review of the 19th century scientific context in which Freud's research lay down its roots. Beyond the physiological and anatomical references of *Project for a Scientific Psychology* (Freud, 1895a), the physiology Freud makes reference to is in reality primarily anchored in an epistemological model derived from physics. Whilst across the Rhine, the autonomy of physiology in relation to physics was far from being accomplished, as a counterpoint, in France, the revolution in physiology driven by Claude Bernard established itself autonomously from physics,. In contrast, Freud's scientific landscape is entirely dominated by the physics elevated to the rank of an ideal science. The influence of Helmholtz, who is both a medical doctor and a physicist, has a determining influence on Freud's training. The discoveries in physics at that time, in particular the formulation of the principle of 'conservation of force' – first principle of thermodynamics – will constitute the points of reference upon which Freud will elaborate his energetics model, then subsequently, the idea of economy in his metapsychology. In this way we can trace both the historic and epistemological path that led Freud from a concept based on physics, and more specifically thermodynamic energy, to an idea of nervous energy that constitutes the basis of the concept of "quantity" as it is stated as 'first fundamental idea' in *Project for a Scientific Psychology* (Freud, 1895a). This notion will subsequently evolve, and lead Freud to the introduction of the concept of 'psychical energy,' this time in a purely metapsychological sense.

Keywords: energy, Freud, thermodynamics, epistemology, Helmholtz, physics

OPEN ACCESS

Edited by:
Jim Hopkins,
University College London,
United Kingdom

Reviewed by:
Michael B. Buchholz,
International Psychoanalytic University
Berlin, Germany
Claudio Colace,
Azienda Sanitaria Locale di Viterbo,
Italy

***Correspondence:**
Jessica Tran The
jessica.tranthe@ens.fr

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *28 June 2018*
Accepted: *11 September 2018*
Published: *11 October 2018*

Citation:
Tran The J, Magistretti P and Ansermet F (2018) The Epistemological Foundations of Freud's Energetics Model. Front. Psychol. 9:1861. doi: 10.3389/fpsyg.2018.01861

INTRODUCTION

The economics point of view of Freudian metapsychology today offers astonishing points of convergence with recent discoveries in contemporary neurosciences, in that it is based explicitly from its first formulations on an energetics model. In 1895 Freud had, in *Project for a Scientific Psychology,* proposed a first principle for the functioning of the psyche, the "principle of neuronal inertia" (Freud, 1895a, p. 296). Freud defined this as the tendency of neurones to divest themselves completely of the quantities of excitation (endogenous or exogenous) that erupt into the psychical apparatus. The primary function of this apparatus would therefore be to reduce to the lowest level possible – ideally a 'level = zero' - the quantity of free energy. This fundamental hypothesis of all the

economic dimension of Freudian metapsychology – which would be later refined in the definitions of the principle of pleasure and the death drive – has recently been put into perspective through the neuroscientific work of Karl Friston and his colleagues. Indeed Friston and his colleagues see cerebral functioning through a Bayesian approach, its aim being to avoid too great a variation in the quantity of free energy coming from our sensorial perceptions (both internal and external) on the basis of prediction of sensorial data. This offers an unexpected point of dialog between psychoanalysis and neuroscience centered on an energetics concept of cerebral function.

While the energetics concept of the psychical function proposed by Freud can offer fruitful points of discussion with neurobiology, it is important to return to the epistemological roots of Freudian energetics in order to pin-point their theoretical origins. The first arguments of *Project for a Scientific Psychology* (Freud, 1895a) do indeed seem to correspond to a biological, even neurobiological, model. In his text, Freud introduces a 'theory of neurons'. This theory constitutes one of the 'two fundamental concepts' that he bases his work on, alongside the concept of 'quantities' understood in terms of energy. However, a brief excursion into the history of the epistemological origins of his concept of nervous energy will enable us to glimpse that this is not truly a biological model. It is on the contrary a paradigm that is radically linked to physics, inspired by work done on the conservation of force, and profoundly influenced by the School of Helmholtz. This journey, following the origins in physicalism of Freudian energetics, will thus serve as a basis for a model dialog between psychoanalysis and neurosciences, where the heterogeneous epistemological roots of these two disciplines can be taken into account.

THE PHYSICALIST PARADIGM OF THE BERLINER PHYSIKALISCHE GESELLSCHAFT

In France, beginning in the 1860s, the "revolution in physiology" (Prochiantz, 1990) brought about by Claude Bernard made a radical epistemological leap by establishing physiology as a discipline in itself. This was autonomous in relation to physics–chemistry (even though Claude Bernard will always postulate a strict physico-chemical determinism of the vital phenomena). Bernard did indeed claim the existence of an undeniable singularity of the *vital aspect* amongst all other physico-chemical aspects, while at the same time strongly criticizing any vitalist stance. According to him, within the organic, "the mechanism is special [. . .], the agent is specific, though the result is identical. No single chemical phenomenon occurs within the body similarly to outside of it" (Bernard, 1885, p. 219). It was in this singularity of the vital mechanism that the concept of homeostasis was rooted, and was later theorized by Cannon, to then be elevated to the rank of physiological mechanism central to all biology.

It was at a time exactly contemporary to the last Bernardian conceptualizations (the *Leçon sur les phénomènes de la vie* are published in 1878, the year Claude Bernard dies), but in a radically different geographic and scientific context, that Freud undertook his medical studies at the Vienna Faculty in the autumn of 1873 (Jones, 1953). Across the Rhine the autonomy of physiology in respect to physics was far from having been realized, and it was on the contrary within an epistemological paradigm that was decidedly antagonistic to that of French biology, that Freud's scientific training was conducted. At the conclusion of his third year Freud joined Ernest Brücke, whom he saw as a "model" (Freud, 1925d, p. 9), at his laboratory of physiology. Besides the respect and admiration Freud felt for this undisputed master (Jones, 1953), this filiation bore witness to an affiliation to a whole scientific paradigm of which Freud will make himself the heir. As Jones underlines, Brücke's institute was closely connected with the school of Helmholtz. The story of this scientific movement had begun in the 1840s with the friendships between different physiologists trained in Johannes Müller's theories on the energy specific to nerves (Assoun, 1981). Du Bois-Reymond, Brücke, Helmholtz, and Ludwig came across as medical doctors imbued with a real "crusading spirit" (Jones, 1953) who, as Du Bois-Reymond reported, had "pledged a solemn oath to put into effect this truth: "No other forces than the common physical-chemical ones are active within the organism." " (Jones, 1953, p. 40). Although they all had medical training, their scientific ideas where totally subordinated to physics. This small group, increased by the addition of new members, young student physicists and physiologists in leagued against vitalism, became in 1845 the *Berliner Physikalische Gesellschaft*, Berlin Physical Society (Jones, 1953). Within less than 30 years they will dominate the German scientific landscape becoming the most influential professors of medicine and physiology of their time, and in turn training a whole generation of students to which Freud and Wundt belonged. The majority of these professors can be equated with what Paul-Laurent Assoun calls the figure of the "doctor-physician," of whom Fechner, Helmholtz, or Lotze will be the principle representatives: "all of them come to physics through medicine or via physiology" (Assoun, 1981, p. 59). For some of them, psychology would constitute the final stage of the journey. It is this scientific practice characterized by its diversity and lack of specialization that Freud would inherit during his years of training at the Brücke Institute. However, it is physics that constitutes for all the related disciplines the epistemological model *par excellence*. It can be observed that the German school of physiology positioned itself in a movement that was the exact opposite of Bernardian physiology: where in France there was a call for a certain independence of physiology as a separate science, autonomous from physics, the Berlin medical practitioners would on the contrary seek to subordinate physiology to physics, making it an extension of the latter. Brücke thus appears as one of the paradigmatic representatives of this trend:

"What is physiology in Brücke's eyes? This is not a pointless question to ask, for the mistake would be to project onto that word the concept formed in the parallel tradition, in France, by Claude Bernard. Physiology for Brücke, leader of the Berlin Physics Society around 1845, is an extension of physics. It has as its object specific physico-Chemical systems, the organisms [. . .]. The physiologist is none other than the physicist of organisms" (Assoun, 1981, pp. 101–102).

THE INFLUENCE OF HELMHOLTZ

Thus it was to a physiology radically subordinated to physics, the overruling dominant science – to which all natural phenomena must be brought back, including those relative to living organisms – that Freud would make himself heir. It is within this orientation that he trained at the Brücke Institute. However the dominant influence of Helmholtz, who of all the scientists at the Berlin Physics Society was without doubt the most eminent, needs to be emphasized. Freud considered him one of his "idols" (Jones, 1953), and would always regret not having had the opportunity to meet him in person. Helmholtz, perfect embodiment of the figure of the 'doctor-physician,' would dominate the German university scene at a time when it was becoming a model and a center of European science. According to him: "all natural phenomena must be brought back to the movement of material particles endowed with invariant driving forces, dependent only on their spatial location" (Prigogine and Stengers, 1979, p. 148). In this he made himself the advocate of an understanding of nature based on mechanical ideas, and the majority of the physiologist of the powerful German school (Liebig, Ludwig, Müller, Du Bois-Reymond, Virchow, Brücke) would adopt his concept according to which "the physico-chemical functioning of the living organism is subject to the same laws as inanimate matter, and must be studies within the same terms" (Prigogine and Stengers, 1979, p. 148).

To understand the influence of this physicalist model in 19th century Germany, it is important to underline that it made its appearance as a reaction to the influence of Schelling's *Naturphilosophie*. A Romantic philosophy that argued for a pantheist monism close to mysticism (Jones, 1953). This philosophy saw nature as a unique fundamental great organism, unified by general laws, by a single principle of causality, and without remainder (Prigogine and Stengers, 1979). While this idealistic concept of nature was spread across all of Europe, German *Naturphilosophie* was characterized by its aspiration to what Schelling described as a 'speculative physics.' It is against this romantic view of a speculative philosophy, for which Helmholtz or Du Bois-Reymond feel a real aversion, that the physicalism of the Berlin Physics Society positioned itself (Meulders, 2001). Freud had been tempted in his youth by these ideas before definitively converting to the views of physicalist science. It was, according to Ernest Jones, under the influence of Goethe that Freud went through a brief period of *Naturphilosophie*, before becoming enthused by the competing physical physiology. In *The Interpretation of Dreams* (Freud, 1900a), Freud mentions that 1 day a violent philosophical discussion with a student, partisan of natural philosophy, nearly led him to a duel. Jones comments on this reactive movement in these terms:

"Physical physiology – although not by itself – overthrew this philosophy and took its place. As has happened before, the conqueror introjected the emotionalism of the victim. 'Unity of science,' 'science,' 'physical forces' were not merely directing ideas or hypotheses of scientific endeavor: they became almost objects of worship. They were more than methods of research – they became a *Weltanschauung*." (Jones, 1953, p.43)

This very strong physicalist scientific ideal, almost raised to a status of religious conviction, appears then to be a virulent reaction to any vitalist views. The radical character of this epistemological model, where philosophy finds itself completely subjugated to physical science, contrasts in a notable way with the Bernardian view which according to Canguilhem constitutes a third pathway between vitalism and reductionism (Canguilhem, 1994). Thus, as Alain Prochiantz points out, up until Claude Bernard "the relationship of biology to physics was divided between complete assimilation in a physicalist reductionism, and radical separation within French vitalism or the German natural philosophy" (Prochiantz, 1990, p. 35). Bernardians therefore rejected both mechanism and vitalism to constitute a third position, adjusting the technique of biological experimentation to the singularity of its object of study, the living.

If the radical physicalism of the German school of physiology is to be understood through the prism of this opposition to the dominant position of natural philosophy (a position whose influence on scientific speculation had been much stronger across the Rhine) it is also to be situated in the context of the discovery in physics of the conservation of energy. Helmholtz was one of the first theoreticians of the conservation of energy. We can, along with Prigogine and Stengers, note the paradoxical fact that the philosophical past of Germany had imbued the scientists, in spite of them, with "an idea far removed from the strictly positivist knowledge that they professed to practice: the idea that nature, in its entirety and without remainder, is unified by a general law, by a single principle of causality" (Prigogine and Stengers, 1979, p. 175). Something akin to a return of the suppressed *Naturphilosophie* against which they had positioned themselves.

THE PRINCIPLE OF THE CONSERVATION OF FORCE

Unifying of physiology and physics resulted from this supposed universal principle of the conservation of energy, according to which "the sum of the forces remains constant in all isolated systems" (Assoun, 1981, p. 102). If Helmholtz was one of the first theoreticians of the principle of the conservation of force, it is Mayer who is considered to have introduced the fundamental distinction between force and matter. Thus it is not surprising, in view of the continuity that existed between physiology and physics to which it is subordinated, that one of the major discoveries of 19th century physics, the principle of the conservation of energy introduced by Mayer in 1852, would have significant consequences for the development of physiology, psychology, and ultimately psychoanalysis (Assoun, 1981). However it is probable that for these two figures of the 'doctor-physicist' that are Hemholtz and Mayer – whose mixed practice was correlative of the porous nature of the boundaries between connected disciplines in the German scientific context – it was originally the studies of living organisms that conferred upon them the intuition of this principle. Their experimental and theoretical scientific practice was in reality so interconnected, moving between the study of living organisms and the inanimate, that it is difficult to establish the physical or physiological

pre-eminence of the observation of the conservation of force (which will subsequently be reformulated as the principle of the conservation of energy). It would be more relevant, rather than to look for such a pre-eminence, to underline that these 'great men' of the 19th century [as Ostwald calls them in his biographical study (Ostwald, 1912)] were accustomed to a mental gymnastics that allowed them to move without difficulty from physiology to physics, and back again. Their experimental practice in one of these two sciences, leading them naturally to theoretical formulations that were equally valid in the other. This practice was justified by a presupposition inherited from Kantian philosophy: nature is ruled by the law of causality, in so far as all change in nature is due to a sufficient cause. As we have seen though, some remanent of the influence of the *Naturphilosophie* (to which they were in fact vigorously opposed) led these doctor-physicists to seek a single principle of causality that would unify nature (both organic and inanimate) in a whole without remainder (Prigogine and Stengers, 1979).

According to Prigogine and Stenger when Mayer, as a young doctor in the Dutch colonies of Java, observed the bright-red color of one of his patients' blood, he concluded from this that since it was warmer in the tropics the inhabitants would need to burn less oxygen. On the basis of this observation, he made an assessment of the consumption of oxygen, which could be considered as a source of energy, and consumptions linked to the maintaining of body temperature with respect to thermic loss, and to manual labor. Mayer generalized the implications of this assessment (which already amounted to an interpretation with respect to the observed facts) to conclude the existence of a "single and indestructible force that is the basis of all phenomena both of living and inanimate nature" (Prigogine and Stengers, 1979, p. 175–176). This thesis of a single energetic principle offered to physiology the grounds for its claim to reduce the 'vital process' down to a mechanical chain of event (Assoun, 1981). Mayer's observations in the tropics thus led him to argue that the body's heat was the result of the chemical energy of food, and he went so far as to assert that the mechanical energy of muscles came from the same origin: mechanical energy, chemical energy, and heat would thus be equivalent and mutually convertible. Back in Germany, he established himself as a doctor and continued his research. He went on to demonstrate that there exists an equivalence between thermic and mechanical work, and calculated that the quantity of heat was equal to a given quantity of mechanical energy. He would present this thesis in 1842 in his *Remarks on the Forces of Inorganic Nature* (Brossollet, 2018).

As Paul-Laurent Assoun points out, "Mayer appears as the Lavoisier of the 19th century, perpetuating the grand principle of the conservation of matter" (Assoun, 1981, p. 60) turning it into a principle of conservation of force; which would become, after the introduction of the terms by Thomson in 1850, the principle of conservation of energy. This discovery could therefore, have come from physiology, to then be theorized and formulated in mathematical equations in the domain of physics, before returning to physiology where its consequences would lead to the development by Wundt of scientific psychology, on the basis of this same principle. Wundt would, it appears, have extended "for the first time, the law of conservation of force to the area of psychology" (Assoun, 1981, p. 60). The discovery of the principle of conservation of energy would also have a notable influence on the thinking of the young Freud. However, this circular phenomenon within German scientific knowledge makes it difficult to establish with certainty the provenance of this principle, whether it was from physiology or physics. Researchers like Mayer or Helmholtz moved into a larger unifying paradigm, where the absence of clear cut lines between disciplines, and a mixed scientific practice (combining both medical practice and research in physics) was completely foreign to the establishing of distinctions between disciplines. This was made possible in France by the advent, with Claude Bernard, of experimental physiology. We can thus argue, as Paul-Laurent Assoun proposes, that from the 1840s "a kind of practice is put in place, that comes simultaneously from physiology, physics, and chemistry; emerging from common and converging interests within a matrix of energetics" (Assoun, 1981, p. 60).

For Mayer then, the *vital aspect* resulted from the transformation of force or matter, and the task of physiology would consist henceforth in the investigation into the mechanisms of this transformation (Assoun, 1981). His work would also be in close relation with the experimental chemistry ushered into Germany by Liebig. Liebig who would contribute to the development of organic chemistry through the study of the chemical processes of living matter (it is indeed in Liebig's review that Mayer's historic memoir on the conservation of force was published). The chemistry of Liebig is essentially analytic: his method consisted in an analysis of the constituent parts of organisms, and he held that it was possible to go from a vegetable compound to an animal compound through the subtraction of constituents (Assoun, 1981). As Paul-Laurent Assoun remarks, this analytic organic chemistry of Liebig's, in close relation with the work of Mayer, would make a deep impression on Freud who, in giving the name of 'psychoanalysis' to his discovery, would borrow specifically the term 'analysis' from the breaking down of the chemical compounds in experimental chemistry inspired by Liebig (Assoun, 1981).

In 1842, in his *Remarks on the Forces of Inorganic Nature,* Mayer did not yet refer to the concept of energy (which will only appear after 1850 in the writings of Thomson, then Rankine), but held to the definition of the concept of 'force.' Thus "the Mayerian project was clearly to ensure the epistemological promotion of the idea of 'force'" (Assoun, 1981, p. 159). While in Germany, the dynamical view inherited from Leibniz and Kant had made force the primary concept, the school of Laplace considered it as an emanation derived from matter. Mayer did not adopt a strict dynamism, since he established an analogy between matter and force. According to him, matter was the fundamental concept of chemistry, ponderable, transformable, and quantitatively indestructible during chemical reactions (where mass would always be conserved, although the quality, for example of oxygen and hydrogen, will not be seen in water) (Locqueneux, 2009). We can already observe here, in chemistry, the hypothesis of a qualitative transformation that nevertheless implies that quantity remains the same; something that Mayer would also apply to force. As Robert Locqueneux underlines, "the role held by matter in chemistry should, according to Mayer, be played by force in

physics" (Locqueneux, 2009, p. 113). Force would therefore also be an indestructible and transformable entity just like matter; but, unlike matter, imponderable. Thus, the inanimate forces of nature could take on multiple qualitative forms: kinetic, thermic, magnetic, electrical, or chemical force. These are qualitative forms that would be phenomenologically distinct manifestation of a same entity, an *Urkraft*, 'elemental force,' (a quest for a primal unique force that is not dissimilar to the *Naturphilosopie's* project, although Mayer was opposed to this). Yet, Mayer was the first to try to establish a quantitative science of force and no longer an only qualitative one: he determined through calculation the mechanical equivalent of heat. The aim was not here to make heat a kind of movement, but to "determine the equivalence between the disappearance of a quantity of heat, and the simultaneous production of movement" (Locqueneux, 2009, p. 114). In this way he would propose an equation for the equivalence between heat and movement (that is to say work in the mechanical meaning of the word). For indeed, according to Mayer,

"If two metals are rubbed together, there is movement that disappears and heat that develops; hence the question: if the movement is the cause of the heat [. . .]. If we cannot account for the disappearance of the movement, without admitting a causal relation between the movement and the heat; it is not possible to understand, without admitting this connection, how the heat develops. It is demonstrated that, in many cases, the disappearance of the movement has no other appreciable consequences than this production of heat" (Mayer, In: Locqueneux, 2009, p. 131).

Mayer then, assumed that the latent heat was transformed in its entirely into a quantity of work: this was a qualitative transformation (the mechanical force was turned into a force of a different nature on a phenomenal level, heat), but one that implied a conservation, an equivalence, from a quantitative point of view. He illustrated this with the example of hydraulic mechanisms where the movement, as it destroys itself, provokes a considerable quantity of heat; and also with steam machines where the reverse takes place, it is the heat that provokes movement. There is then a principle of equivalence, that also implies a reversibility, between movement and heat, as two qualitatively distinct aspects of a same original force (Locqueneux, 2009).

Three years after this work on the forces in inanimate nature, Mayer would return to questions of physiology and write a memoir on *The Motions of Organisms and their Relation to Metabolism* (1845). He described his project as a desire to "fill the chasm that separates exact physics and physiology" for which "a method that would seek to bring together these two sciences under this one perspective, would be invaluable for physiology" (Mayer, 1845, In: Assoun, 1981, p. 163). This union between physiology and physics was thus sealed by the reunion of the heterogeneous phenomena observed by these two disciplines under an overruling principle. This overruling principle would be the conservation of force – subsequently translated into conservation of energy. In both organic and inorganic phenomena, there would therefore be only one force at work. A force that would manifest itself under a qualitatively distinct phenomenology: "this circular force through a perpetual exchange, in inanimate nature as well as in living nature. In both domains, there is no phenomenon without transformation of force" (Mayer, 1845, In: Assoun, 1981, p. 164), that force remaining constant beyond all its transformations.

If it has therefore been so essential to consider, at such length, the concept in physics of the principle of conservation, it is that, as has so rightly pointed out Paul-Laurent Assoun, "not only do physiology and physics take their inspiration from it, they are also closely involved in its evolution. It is too little to say that physics extends to or is applied to psychophysiology, there is an imbrication of the two" (Assoun, 1981, p. 166). We have been able to observe that the intuition of the conservation of energy had been for Mayer influenced by his medical practice, in connection with his physiological research on 'body heat.' It was in effect while meditating on the production of body heat through combustion, that he was able to deduce the principle of conservation. However, after having theorized from a purely physical point of view the conservation of force in 'inanimate nature,' he would come back to a physiological application of this energy gain. Thus, "energetism is introduced into psychophysiology not by a simple extension, but as an annex field of verification of one and the same idea" (Assoun, 1981, p. 166) The implication of this is that at no time did Freud feel that he was 'borrowing' concepts from physics or physiology, rather he was 'managing his property,' in as much as this energetics model was an inherent part of his 'scientific cradle.' In this sense, *Project for a Scientific Psychology* (Freud, 1895a) is a paradigmatic example of the importance of this heritage (Assoun, 1981).

While Mayer was the first to establish the principle of conservation of force in physics, and to give an equation for the equivalence between heat and movement, Helmholtz would pursue his project by applying this principle to physiology (Assoun, 1981). Freud would recognize him as his idol – indeed Helmholtz would dominate the German University scene in the 19th century. It was then, essentially through Helmholtz's work, that Freud would assimilate the principle of conservation of energy, and its applications to physiology. In his memoir of 1847, *On the Conservation of Force,* Helmholtz in the first instance excludes any possibility of 'perpetual motion'. In this he was going against the generally held view according to which, an inexhaustible and constantly renewed 'vital force' would maintain the activities of living organism, and would control the activities of physical and chemical forces (Locqueneux, 2009). The impossibility of perpetual motion had already been justified in mechanics. However, Helmholtz extended it to the whole of nature by applying it to natural forces of a different order to mechanical forces, forces such as heat, electricity, magnetism, light, and chemical reactions: "there does not exist, in all the series of natural actions, a process that would permit the creation of mechanical force without an equal expenditure" (Helmholtz, 1847, In: Locqueneux, 2009, p. 124). Thus it would be impossible to imagine, in nature, a machine that would present perpetual motion – a swing or a pendulum, for example, would eventually stop under the effect of friction and the loss of heat that results (Meulders, 2001). We might be tempted to counter this impossibility with the definition of inertia, posed by Newton as

the first law of classical physics, that consist in the tendency of bodies to maintain their speed. However, this universal property only allows for the conceptualization of perpetual motion under abstract conditions, in a closed system, protected from any other force other than the one that had caused the body's speed. This case is thus utopian, and not observable in nature. It would require the existence of a closed space, in a vacuum, and without friction. Yet, even in this fictional case it would be, as Michel Meulders points out, inappropriate to speak of 'perpetual motion,' since at rest bodies are immobile and that only an external force could have put them into motion (Meulders, 2001). There exists then a certain ambiguity, as Pierre Costabel points out, in classical physics', between its posit of the impossibility of perpetual motion and its definition of inertial movement (Costabel). For Helmholtz then, the impossibility of perpetual motion was correlative to the principle of conservation of force, according to which there could only be creation of movement through a corresponding expenditure of energy. Thus the impossibility of perpetual motion is heir to Leibniz's dynamics. Leibniz had indeed been the first to profess the impossibility of 'mechanical' perpetual motion based on the elementary metaphysical principle that it is impossible to create from nothing, *ex nihilo* (Costabel).

For Helmholtz, all actions of nature must then be brought back, in the final instance, to the opposition of the two forces of repulsion and attraction, as they were formulated by Newton: "thus the problem for the physical sciences consists in bringing all natural phenomena back to invariable forces, attraction and repulsion, whose intensity depends on the distance from the centers of action" (Meulders, 2001, p. 128). On the basis of these two presuppositions (the impossibility of perpetual motion, and reduction of all the actions of nature to the forces of repulsion and attraction), Helmholtz came to establish, in his memoir, the principle of conservation of 'vital force.' The problem he was confronted with, and which led him to the definition of this principle, can be summarized by the paradigmatic example of the movement of a swing, which, when it reaches the highest point of its movement, finds itself for a brief moment immobile, before beginning its descent under the influence of gravity. At that point it once again gains speed and goes up in the opposite direction, fighting gravity, before decelerating, and stopping once again under the effect of gravity. The appearance is then that there are two forces involved: one caused by gravity, and the other operating in an opposite direction through the effect of the speed gained by the swing. The first force would be at its maximum when the swing is at the highest point, whilst the other force would reach its paroxysm when the swing goes by, very fast, on the vertical. According to Michel Meulders, "everything seems to the 'naive' observer as if these two 'forces' to have a mysterious relationship to each other, in which the increase in one would lead to the decrease of the other, and vice versa" (Meulders, 2001, p. 129).

Alongside the 'live force,' which here corresponds to the effective movement of the swing, Helmholtz established the necessity to introduce a 'tension force' that corresponded to the potentiality of movement to come, when the swing is at rest at its apex. The 'tension forces' of material bodies were defined as the product of the forces of attraction or repulsion and the distance that separates those particles. The 'live forces' were themselves linked to the movement of these particles (Locqueneux, 2009). The dynamics of Helmholtz was thus entirely reducible to a mechanical description of nature. In this way he referred all the forces present in nature (for which Mayer had compiled a list, as we have seen), examples being electrical, chemical, live and calorific force, back to these mechanical forces that were 'live force' and 'tension force.' He then referred to the results established by Joules between the loss of mechanical force and the giving off of a quantity of heat (for instance emitted by friction). These results had given a mathematical formula for calculating the rise in temperature (the number of degrees of elevation) in relation to the friction (produced here by the elevation of a weight), which Mayer had been the first to formulate.

By using this quantitative mathematical result for the equivalence between heat and mechanical movement, Helmholtz contended that "the quantity of heat can be increased in an absolute manner by the mechanical forces" (Helmholtz, 1847, In: Locqueneux, 2009, p. 126). From a resolutely mechanistic stand point he then reduced heat to a quantity of movement. He argued that what had up until then been called 'quantity of heat' would in fact only be another way of expressing a 'quantity of live force' of movement within a substance, as well as the 'quantity of tension force' of the internal state of that substance. The first would correspond to free or perceptible heat, whilst the second would correspond the latent heat (Locqueneux, 2009). Heat then, would demonstrate the same distribution between potential forces and their active expression (if we refer to the Aristotelian model), the tension forces and 'live forces.' Helmholtz concluded from this that all natural phenomena, whether they applied to organic or inanimate substances, were caused solely by 'live force's (*veres vivae*) and tension forces (*Spannkraft*). The concept of force, in a Kantian perspective, thus allowed nature to be rendered intelligible, and to unify the knowledge we have of it (Locqueneux, 2009).

THE ENERGETICS MODEL

Mayer and Helmholtz made the principle of conservation of force a fundamental principle, one that allowed for the union between physics and physiology. However, the concept of energy does not yet appear in Mayer's memoir, nor in Helmholtz's *On the Conservation of Force*, published in 1847. It was only with the introduction of the concept of energy by Thomson, then Rankine, that Helmholtz adopted this new terminology. Helmoltz then rewrote his concepts of 'live force' and 'tension force' in terms of kinetic energy and potential energy. Terms that would then be taken up by Breuer in his *Studies on Hysteria* (Freud and Breuer, 1895b), and will thus have a notable influence on Freudian theory.

Thomson mentions, for the first time, the word 'energy' in 1850. He retained the word 'force' to designate Newtonian forces defined by the laws of movement. Energy, would thence designate all the other kinds of force that Mayer had enumerated in his memoir. Thomson considered the existence of two categories of energy, a static energy and a dynamic energy (Locqueneux, 2009). These could also cover the distinction between a latent

force and an active force that was already present in Helmholtz's definition of 'live forces' and tension forces. Thus, in his paper *On the universal Tendency in Nature to the Dissipation of Mechanical Energy*. (Thomson, 1852), Thomson states that:

"a load suspended and ready to fall, an electrified body, a quantity of fuel or coal, contain reserves of energy of a static nature, a physical body in motion, an area of space crossed by light or radiant heat waves, a body whose molecules are agitated, contain reserves of energy of a dynamic nature" (Thomson, 1852, In: Locqueneux, 2009, p.127)

It needs to be pointed out that only measurable physical quantities were mentioned here: indeed, the question of the quantification of energy would remain one of the most important objectives for all 19th century German physics. This in opposition to *Naturphilosophie*'s purely qualitative terms for the description of nature.

It was also in this same paper that Thomson expressed for the first time the second principle of thermodynamics. After having regrouped everything that Mayer designated with the term of force (mechanical, chemical, magnetic, calorific. . .) under the one concept of energy; he turned his attention to the output of real machines, that demonstrated a loss of energy – whereas Sadi Carnot had laid down the grounds for a principle of conservation, working from an abstraction of ideal machines which would experience no reduction in yield. In this way, Thomson concluded that when heat passed by conduction from one body, to another body at a lower temperature, some wastage of mechanical energy would occur. This loss of mechanical energy, through thermal conduction, would have as a consequence, an irreversibility in the processes taking place in thermodynamic machines. This observation would then be generalized to the statement of a continuous degradation of energy in the universe (Locqueneux, 2009). Thus, the irreversible propagation of heat – synonymous in the context of thermodynamic machines with a loss of yield – would become, from 1852 onward, the tendency to the universal degradation of mechanical energy (Prigogine and Stengers, 1979). As Prigogine and Stenger point out: "In this way, Thomson makes the vertiginous leap from the technology of motors, to cosmology [. . .]. Thomson's new theory [. . .] also makes manifest the consequences of the irreversible propagation of heat in a world where energy is preserved, this world [. . .] can only be at the cost of an irreversible waste, a useless dissipation of a certain quantity of heat. The differences that produce effects are ceaselessly diminishing within nature" (Prigogine and Stengers, 1979, pp. 184–185).

From the 1850s onward, we find the two principles of thermodynamics formulated almost in their definitive forms: the conservation of energy, and the entropy principle. The reformulation by Thomson, of Helmholtz and Mayer's work on the conservation of force, marked the consecration of the energetics model, and would henceforth dominate the German scientific landscape of the second half of the 19th century. In 1853, Rankine performed some level of synthesis of the Helmoltzian distinction between the 'live forces' and the tension forces, and the unifying concept of energy introduced by Thomson, by separating energy into two categories: potential energy (contained in material constructs capable of producing work), and kinetic energy. This separation was applied not only to mechanical force, but also to all kinds of physical phenomena; and covered the Aristotelian concepts of *dynamis* and *energeia*, potency and actuality. In an article written in 1862, Thomson would substitute the term kinetic energy for that of actual energy, and it was this terminology that Helmholtz would use when he went on to adopt the term of energy rather than that of force (Locqueneux, 2009).

Henceforth Helmholtz would favor the concept of energy over that of force, in so far as although the latter remains the ultimate cause of movement, energy as a quantitative concept, measures the capacity of a system to realize, under the impulsion of a force, a certain quantity of work, be it mechanical, caloric, chemical or electrical (Meulders, 2001). To go back to the example of the swing that had illustrated the difference between 'live force' and tension force: energy, that is to say the capacity of the swing to perform, under the impulsion of a force, its mechanical pendulum movement, can be considered as the sum of two energies, one kinetic (that of active movement) and the other potential (containing the latent movement). The first would consist of the speed of the given movement, whereas the second would be relative to the position of the swing in space, subject to the forces of gravitation and gravity (Meulders, 2001). Helmholtz held that the sum of the kinetic and potential energies always remained constant: "In all instances of the movement of free material points under the influence of their forces of attraction or repulsion, the intensity of which depends only on distance, the reduction of potential energy is always equal to the increase in live force (kinetic energy). The sum of the live forces and of the potential energy is always constant" (Helmholtz, 1882a,b, In: Meulders, 2001, p. 130).

This integration by Helmholtz of the concept of energy to his previous developments on the principle of conservation of force, and the distinction between 'live force' (which will become kinetic energy) and 'tension force' (that will subsequently called potential energy), bears witness to the resolutely mechanical character of his references to energetics. Helmholtz, along with Joule or Rankine, used the concept of energy to extend mechanical principles to other non-mechanical domains. This was contrary to Mayer, who saw mechanical phenomena as simply consisting in a particular instance of the phenomena of transformation of energy (Assoun, 1981). It is right to underline that Ostwald, professor at Leipzig since 1887, would specifically oppose this 'energetic mechanism' or 'mitigated energetics,' raising energetics to the rank of doctrine, assimilable to a quasi-theological *Weltaschaung*, and very close to *Naturphilosophie*. He would propose a new philosophy of science, the fundamental concept of which would be that of energy (Assoun, 1981). His integral energetics formed the basis for an immaterial ontology, and was akin to a radical monism where "nothing seems to be able to occur without energy being a part of it" (Ostwald, 1891, In: Assoun, 1981, p. 170), whereas Mayer had held on to a dualist model, considering matter and force as two distinct entities. Ostwald put himself in opposition to the definition of a potential energy (inherited from the mechanical concept of tension force), that would

erase the original and universal reality that is energy in its actuality.

Freud, as we have seen, idolized Helmholtz. Furthermore, he would take as his reference the definition, inspired by Helmholtz, and put forward by Breuer in *Studies on Hysteria* (Freud and Breuer, 1895b), that postulated the existence of a nervous energy, defined as "intracerebral tonic excitation," the nature of which could be quiescent (that is potential) or kinetic (actual) – although Freud would make some alterations to this definition. Freud's model then was resolutely that of the mitigated energetics of Helmholtz, in so far as in essence he used a functional energetics applied to the functioning of the psyche, and would regularly use the term 'work' to describe the processes of the unconscious (dream work, work of mourning, etc.). His description of the movement between different psychological states, that would entail a mechanical expenditure, would also be a "specific expression of the general rise in disorder that the second principle of thermodynamics formulates" (Assoun, 1981, p. 182). This is why, as Paul-Laurent Assoun points out: "Freud never encounters the temptation, inherent to doctrinal energetics, to exalt energy as a supra-mechanic active principle, and to hypostasize it in support of a world view" (Assoun, 1981, p. 182). Energetics will constitute the basis of all the economic aspect of the metapsychology, but "never will this model of deciphering hypostasize into an energetics doctrine" (Assoun, 1981, p. 182–183).

FROM NERVOUS ENERGY TO PSYCHIC ENERGY

The introduction in the 1850s of the concept of energy, propelled notably by the influence of Mayer and Helmholtz's work, appears as the heuristic key for the coming together of physiology and physics. This laid down the foundations, as we have seen, for a resolutely physical epistemological model, which Freud in turn wholeheartedly adheres to. This model positioned itself in a radically autonomous position in relation to the Bernardian revolution in France, which had promoted the singularity and the independence of physiology as a separate science. From then on, Helmholtz, and in his wake Brücke, would give themselves the task of applying the physical principle of the conservation of energy, to organic phenomena. The concept of energy would thus make it possible to simultaneously encompass the concept of force (kinetic, thermic, electrical, magnetic), and phenomena that belonged to living organisms, such as innervation, irritability, and some chemical reactions (Assoun, 1981).

What Helmholtz would aim to achieve in his work *On the Conservation of Force* (Helmholtz, 1847), consisted in applying the concept in physics of conservation of energy, to biology, by making it a postulate for physiological events (Assoun, 1981). The publication of this work marks an essential turning point in accomplishing the unification of the natural sciences, through the application of the principle of conservation of energy. Helmholz can thus be incontrovertibly recognized as the scientist who opened the royal road for an energetics concept of physiology, as well as of psychology. Thus, as Assoun emphasizes, "When in 1883, Freud admits his idolization of the great Berlin master, he is expressing an emotional adherence to a model that confirms his epistemological position. It is, furthermore, with the man who clinched the union of psychology and neurology that he throws in his enthusiastic lot" (Assoun, 1981, p. 158). In this resolutely Helmholtzian filiation, when it comes to a desire to apply the principles of energetics to physiology and anatomy, we can more specifically postulate the probable influence on Freud of Helmholtz's work on neurons and the speed of propagation of the nervous influx. An area Freud was introduced to when he himself worked on the dissection and observation of nerve cells at the Brücke Institute. Nervous innervation and irritability can thus appear as the energetic manifestations specific to the nervous systems of living organisms.

We have seen that, from an epistemological point of view, Freud did not follow Ostwald's immaterial ontology, which preached an integral energetics assimilable to a *Weltanschauung*. Rather he positioned himself within a mitigated and a mechanic energetics, in a Helmhotzian filiation. Nevertheless, it is right to recognize that Ostwald himself had naturally come to question the possibility for applying the notion of energy to psychological phenomena, to the extent that, in his manifest on *Energy,* he considered the phenomenon of life as a "constant manifestation of energy" (Ostwald, 1891, In: Assoun, 1981, p. 170). Faithful to his pan-energetics view of nature, Ostwald was led to state that "psychological phenomena can be construed as energetic phenomena, and interpreted as such just as well as any other phenomena" (Ostwald, 1891, In: Assoun, 1981, p. 172). Ostwald then, argued for the existence of a nervous energy, and described a process during psychological activity that gives rise to a consumption of energy. In so far as he introduced the concept of 'psychic energy,' he can be said to open up the way for Breuer and Freud's studies on the energy of the nervous system, or "intracerebral tonic excitation" (Freud and Breuer, 1895b) for which Freud would give the term 'quantity' in his *Project for a Scientific Psychology* (Freud, 1895a). Ostwald would thus have been the first to see psychic phenomena as being "phenomena of nervous energy." Going on to define them as being, like all energy, a "measurable quantity that obeys the law of conservation and that of transformation" (Ostwald, 1891, In: Assoun, 1981, p. 172), that could manifest themselves under diverse forms. This need for quantification and measurement of psychic energy will continue throughout the work of Freud. Freud would, however, renounce giving it an absolute value, and accommodate himself to attributing to it a measure that was only relative. The principle of conservation of energy was thus applied by Ostwald to psychological phenomena in so far as, according to him "no psychological operation takes place without a corresponding consumption of energy" (Ostwald, 1891, In: Assoun, 1981, p. 172). Freud would reapply this introduction of the principle of conservation of energy in the field of psychology, but without following the metaphysical consequences that Ostwald upheld (that is to say the circumventing of the "religious problem of the soul and the body" through the concept of psychic energy).

It is then, these models for the implementation of energetics in physiology, and *a fortiori* psychology, that Freud, in the wake

of Helmholtz, would pursue. This was not, however, without some influence from Ostwald; although he in no way adhered to Ostwald's ontological monism, and would retain all through his work a mechanical energetics close to that of Helmholtz. This physicalist paradigm would be enriched in Freud's thinking by his training in anatomy at the Brücke Institute, where he devoted himself to the meticulous study of nerve cells. As Ernest Jones recounts, Freud, then a medical student, was seduced by the psycho-physiological theories held at the Brücke Institute – Brücke supported Helmholtz's school, and also played an important role at his side in the Berlin Physical Society (Jones, 1953). Du Bois-Reymond reports that Helmholtz and Brücke had "pledged a solemn oath to put into effect this truth: "No other forces than the common physical-chemical ones are active within the organism." " (Jones, 1953, p.40). The training that Freud received at the Brücke Institute was, then, dominated by the application of the principle of conservation of energy to organisms: "Organisms differ from dead material entities in action – machines – in possessing the faculty of assimilation, but they are all phenomena of the physical world; systems of atoms, moved by forces, according to the principle of the conservation of energy discovered by Robert Mayer in 1842, neglected for 20 years, and then popularized by Helmholtz. The sum of forces (motive forces and potential forces) remains constant in every isolated system. The real causes are symbolized in science by the word 'force'." (Jones, 1953, p.41).

Here is then, summed up by Jones, the message of the German school of physiology at the heart of which Freud will be immersed during his years of training. The works of Brücke devoted to transformation and to the effects of physical forces in the living organism, will have a lasting influence on the dynamic view of metapsychology; and Freud, up until 1926, would argue that within the psychical apparatus "The Forces assist or inhibit one another, combine with one another, enter into compromises with one another, etc." (Jones, 1953, p. 42).

CONCLUSION

The roots of the Freudian energetics model, heavily influenced by the discovery of the principle of conservation of force in thermodynamics, prompt us to understand to origins of the theoretical model that he develops starting with in *Project for a Scientific Psychology* (Freud, 1895a), not as a primarily biological model, but on the contrary as a paradigm radically grounded in physics. Freud's developments on the principle of inertia, then on the pleasure principle, and later, on the death drive, should then, be re-situated within the scientific project of the Helmholtz school, a project to subordinate physiology, and subsequently psychology, to an ideal of physics. Not disregarding the fact that, in the 19th century German and Austrian scientific context, biology had not yet acquired its independence from the ideal of physics; the model put forward in *Project for a Scientific Psychology* (1895a) should therefore not be too swiftly described as exclusively biological. The physiology training that Freud received at the Brücke Institute is in no way comparable to the Bernardian physiology. Bernard had gone in a different direction introducing a new position. While rejecting vitalism, Claude Bernard admitted the existence of a singularity of the 'vital force,' allowing for an autonomy of experimental physiology as an independent science in relation to the science of physics.

It is within a biology that is firmly subordinated to the *Berliner Physikalische Gesellschaft's* ideal of physics, dominated by the influence of Helmholtz, that the energetics model of the *Project for a Scientific Psychology* (Freud, 1895a) is rooted. These considerations could open up a field of enquiry that would benefit from further investigation: If Freud, notably in the wake of the publication of The Interpretation *of Dreams* (Freud, 1900a), renounces - at least temporarily – grounding his theorizing on a biological model, is this move also accompanied by a renouncement of the physicalist epistemological foundations of that model? This epistemological turning point, already announced in the letter to Fliess dated 6th December 1896 (Freud, 1950 [1892–1899]), seems to represent a turning point in the move from a "neuronal apparatus" model in the *Project for a Scientific Psychology* (Freud, 1895a), to the abandonment of this project for basing mental processes on a precise description of the nervous system. In *The Interpretation of Dreams* (Freud, 1900a) Freud will have indeed abandoned the vocabulary of physiology, no longer referring to the structure and anatomy of neurons, and will henceforth refer exclusively to a "psychical" apparatus. In chapter VII, he formulates this renouncement of an anatomically localized model thus:

"I shall entirely disregard the fact that the mental apparatus with which we are here concerned is also known to us in the form of an anatomical preparation, and I shall carefully avoid the temptation to determine psychical locality in any anatomical fashion." (Freud, 1900a, p. 536)

This is only a temporary renouncement, in so much as that the hope for a physiological model is deferred to such a time as progress in biology will make it possible to precisely base psychical processes on a physical substrate. Nevertheless, this research had henceforth become secondary for Freud, and no longer constitutes the primary aim of his theorization: "We may, I think, dismiss the possibility of giving the phrase an anatomical interpretation..." (Freud, 1900a, pp. 48–49)

Thus, although reference to anatomy is not totally absent from Freud's thinking after 1899, it remains a fact that this search for an anatomical location is sidelined from a topical point of view from the metapsychology. In many ways it is no longer necessary to metapsychology, which can do without it. One question does remain: Is this putting aside (however, temporary it may be) of all references to an anatomical location for the psychical processes accompanied by a renunciation of the physics model? It was this physical model that had enabled him, progressively, to draw out, on the basis of the idea of nervous energy, the concept of psychical energy, that will subsequently become libido. We could be tempted to argue that, despite abandoning a biological reference in the construction of Freudian metapsychology, the influence of a physics epistemological model seems to remain. However, this question is no the object of this research, and will be the focus of work to come.

These historical elements can, then, prompt us to reconsider the major metapsychological concepts through the prism of

the energetics model. In particular the important principles of thermodynamics as they were understood when they came to light at the end of the 19th century. A further in-depth look at the Freudian formulations of the principle of inertia, the pleasure principle and even the death drive, will be enriched by an analysis that seeks to bring to light the links between these concepts and the influence of the physics model, more particularly the model of thermodynamics at the roots of Freudian thinking.

AUTHOR CONTRIBUTIONS

JT is the main contributor of this paper as part of her Ph.D. thesis. FA and PM as supervisors, contributed to the conception and development of the research, and they revised critically the manuscript for intellectual content.

FUNDING

This work was funded by a grant from the Agalma Foundation, Geneva, Switzerland.

ACKNOWLEDGMENTS

We thank the Agalma Foundation for financial support, and Kirsten Ellerby for the translation.

REFERENCES

Assoun, P. L. (1981). *Introduction à l'épistémologie freudienne*. Paris: Payot.

Bernard, C. (1885). *Leçons sur les phénomènes de la vie communs aux animaux et aux végétaux*. Paris: Baillère et fils.

Brossollet, J. (2018). *Mayer Julius Robert Von - (1814-1878). Encyclopædia Universalis*. Available at: http://www.universalis-edu.com/encyclopedie/julius-robert-von-mayer/

Canguilhem, G. (1994). *Etudes d'histoire et de philosophie des sciences*. Paris: Vrin.

Freud, S. (1895a). *A Project for a Scientific Psychology. S.E., 1*. London: Hogarth, 283–397.

Freud, S. (1900a). *The Interpretation of Dreams. S.E., 4*. London: Hogarth.

Freud, S. (1925d). *An Autobiographical Study. S.E.,* 20. London: Hogarth, 3–70.

Freud, S., and Breuer, J. (1895b). *Studies on Hysteria. S.E., 2*. London: Hogarth.

Freud, S. (1950 [1892–1899]). *Extracts from the Fliess Papers. S.E. 1*. London: Hogarth, 175–279.

Helmholtz, H. (1847). *Über die Erhaltung der Kraft, eine Physikalische Abhandlung*. Berlin: Reimer.

Helmholtz, H. (1882a). "On the thermodynamics of chemical processes," in *Physical Memoirs Selected and Translated from Foreign Sources*, Vol. 1 (London: Taylor & Francis), 43–97.

Helmholtz, H. (1882b). Zur Thermodynamik chemischer Vorgänge. *Berliner Berichte* 1882, 22–39, 825–836; 1883, 647–665.

Jones, E. (1953). *The Life and Works of Sigmund Freud*, Vol. 1. New York, NY: Basic Books.

Locqueneux, R. (2009). *Histoire de la thermodynamique classique : de Sadi Carnot à Gibbs*. Paris: Belin.

Meulders, M. (2001). *Helmholtz, des lumières aux neurosciences*. Paris: Odile Jacob.

Ostwald, W. (1891). Studien zur Energetik I. *Berichte über die Verhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig.* 43, 271–288.

Ostwald, W. (1912). *Les Grands hommes*. Paris: Flammarion.

Prigogine, I., and Stengers, I. (1979). *La Nouvelle alliance*. Paris: Gallimard, Folio essais.

Prochiantz, A. (1990). *Claude Bernard, la révolution physiologique*. Paris: P.U.F.

Thomson, W. (1852). On a universal tendency in nature to the dissipation of mechanical energy. *Philos. Mag. J. Sci.* 4, 304–306. doi: 10.1080/14786445208647126

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 Psychology

HYPOTHESIS AND THEORY
published: 30 January 2019
doi: 10.3389/fpsyg.2018.02714

The Hard Problem of Consciousness and the Free Energy Principle

*Mark Solms**

Department of Psychology, University of Cape Town, Cape Town, South Africa

This article applies the free energy principle to the hard problem of consciousness. After clarifying some philosophical issues concerning functionalism, it identifies the elemental form of consciousness as *affect* and locates its physiological mechanism (an extended form of homeostasis) in the upper brainstem. This mechanism is then formalized in terms of free energy minimization (in unpredicted contexts) where decreases and increases in expected uncertainty are felt as pleasure and unpleasure, respectively. Emphasis is placed on the reasons why such existential imperatives *feel like something* to and for an organism.

Keywords: hard problem, consciousness, free energy, predictive processing, affect, Freud

OPEN ACCESS

Edited by:
Andrea Clarici,
University of Trieste, Italy

Reviewed by:
Daniela Flores Mosri,
Universidad Intercontinental, Mexico
Mark John James Edwards,
Institute of Neurology, University College London, United Kingdom

***Correspondence:**
Mark Solms
mark.solms@uct.ac.za

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *03 July 2018*
Accepted: *17 December 2018*
Published: *30 January 2019*

Citation:
Solms M (2019) The Hard Problem of Consciousness and the Free Energy Principle. Front. Psychol. 9:2714. doi: 10.3389/fpsyg.2018.02714

I recently published a dense article on this topic (Solms and Friston, 2018)—a sort of preliminary communication—which I would like to expand upon here, in advance of a book-length treatment to be published under the title *Consciousness Itself* (Solms, in press). Since this is a psychoanalytic journal, I will supplement my argument with cross-references to Freud's views on these themes. Readers with a mathematical background will benefit from a close reading of Solms and Friston (2018) in conjunction with this paper, which is aimed primarily at a psychologically educated readership.

My argument unfolds over four sections, of unequal length. The first addresses some philosophical issues pertaining to dual-aspect monism in relation to the hard problem. The second reconsiders the anatomical localization of consciousness (the so-called neural correlate of consciousness or NCC) in the cerebral cortex. In consequence, it reconceptualizes the functional roles of the "level" vs. "contents" of consciousness. The third and most important section explains the dual aspects of consciousness (its physiological and psychological manifestations) in formal mechanistic terms, in relation to the imperatives of free energy minimization. The fourth section briefly pursues some implications of this formulation for the cognitive neuroscience of consciousness, in relation to memory consolidation and reconsolidation.

THE PROBLEM WITH THE HARD PROBLEM

Does the Brain Produce the Mind?

The original statement of the hard problem, as formulated by David Chalmers, is put like this:

> It is undeniable that some organisms are subjects of experience. But the question of how it is that these systems are subjects of experience is perplexing. Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does (Chalmers, 1995).

A shorter statement of the problem goes like this: "How and why do neurophysiological activities *produce* the "experience of consciousness?" (Chalmers, 1996, emphasis added). John Searle says something similar: "How exactly do neurobiological processes in the brain *cause* consciousness?" (Searle, 2017, p. xiii, emphasis).

The starting point of the argument I shall set out here is that the brain does not "produce" or "cause" consciousness. Formulating the relationship between the brain and the mind in causal terms makes the hard problem harder than it needs to be. The brain does not produce consciousness in the sense that the liver produces bile, and physiological processes do not cause—or become or turn into—mental experiences through some curious metaphysical transformation.

When I wake up in the morning and experience myself (my mind) to exist, and then confirm in the mirror that I (my body) do indeed exist, I am simply realizing the same thing from two different *observational perspectives* (first-person and second-person perspectives). Asking how my body produces my mental experience is like asking how lightning causes thunder.

This is the dual-apect monist position on the mind/body problem[1]. There can of course be no question of determining a "correct" metaphysical starting point, but the dual-aspect monist position—which is the starting point of my argument—raises an interesting philosophical question. If body and mind are two appearances (aspects) of the same underlying thing, then what stuff is the underlying thing made of? In other words, using the analogy of thunder and lightning, what is the metapsychological[2] equivalent of "electricity" (i.e., the thing that gives rise to thunder and lightning, both)?

This question requires one to clarify what we mean (ontologically) by terms like "physical basis," "physical processing," "neurophysiological activities," and "neurobiological processes"—terms which turn out to be surprisingly ambiguous. If physiological phenomena—like their mental correlates—are appearances, then their basis must be something non-physiological.

Let us approach the question by way of an example. If the internal experience of having a memory and the neuronal assemblage embodying that same memory (pictured externally, through optogenetics, for example) are two realizations of a single underlying thing, then what is "memory" itself made of? The answer is that it is abstracted from both manifestations. Memory is not a stuff; it is a *function*. We describe functions in terms of their underlying lawful mechanics, not their appearances[3]. The laws are inferred from the regularities we observe; they *explain* the appearances.

There are of course both psychological and physiological accounts of the functions of memory; but the mechanism a dual-aspect monist is looking for must be sufficiently deep to account equally for both of its observable manifestations—psychological and physiological. In the above example: if we explain the experience of remembering in psychological terms and the activation of the neuronal assemblage (and associated cellular processes) in physiological terms, then our functional inferences are too superficial, and an "explanatory gap" will appear between them (Levine, 1983). Accordingly, one must infer laws which are abstracted equally from the two phenomenal surfaces, sufficiently deeply to underpin the psychological *and* physiological accounts[4].

This is not difficult to do. Consider, for example, short-term memory (STM). Miller's law states that human beings are capable of holding seven-plus-or-minus-two units of information in working memory at any one point in time. This is an abstraction derived both from the (psychological) experience of trying to hold more than seven things in mind and from observing the correlated (physiological) synaptic dynamics of STM traces (Mongillo et al., 2008). The same applies to Ribot's law, concerning the temporal gradient of long-term memory (LTM), which underpins both the psychological and physiological phenomena of memory consolidation over time (Kandel et al., 2012). These laws concern the behavior of an abstracted function, which is (in itself) both psychological and physiological. Ultimately, in all sciences, we aspire to reduce such laws to formalized algorithms—to mathematics—the ideal of third-person abstraction[5].

That is why terms like "physical basis" and "neurobiological processes," etc., are surprisingly ambiguous in relation to mental functions. They suggest asymmetrical (i.e., overly superficial) functional concepts which can explain only the neurological side of the neuro/psychological equation—thereby leaving an explanatory gap.

[1]Freud was a dual-aspect monist (see Solms, 1997). Here, I am disregarding the clinical complexities arising from the developmental achievement of recognizing oneself in the mirror.

[2]When Freud first introduced this term (Letter to Fliess of March 10, 1898; Freud, 1950 [1892-99]) he said it refers to a level of explanation that incorporates both psychology and biology. In this way he aspired to "transform metaphysics into metapsychology" (Freud, 1901, p. 259).

[3]Freud's priority in formulating this "functionalist" position is not recognized: "[We] attempt to make the complications of mental functioning intelligible by *dissecting the function* and assigning its different constituents to different component parts of the apparatus. So far as I know, the experiment has not hitherto been made of using this method of dissection in order to investigate the way in which the mental instrument is put together, and I can see no harm in it." (Freud, 1900, p. 536, emphasis added).

[4]Freud put it like this: "We should picture the instrument which carries out our mental functions as resembling a compound microscope or photographic apparatus, or something of the kind. On that basis, psychical locality will correspond to a point inside the apparatus at which one of the preliminary stages of an image comes into being. In the microscope and telescope, as we know, these occur at ideal points, regions in which no tangible component of the instrument is situated" (Freud, 1900, p. 536).

[5]This was the goal of the Helmholtz school of medicine: "Brücke and I pledged a solemn oath to put into effect this truth: 'No other forces than the common physical and chemical ones are active within the organism. In those cases which cannot currently be explained by these forces one has either to find the specific way or form of their action by means of the physical-mathematical method or to assume new forces equal in dignity'" (Du Bois-Reymond, 1842; Letter to Hallmann, in Du Bois-Reymond, 1918). The young Freud was a pupil of the Helmholtz school, and described Brücke as one of his formative role-models (Freud, 1925a).

But before one can identify the functional laws underpinning the regularities of both conscious experience *and* its neural correlates, one faces a further hurdle.

Is Consciousness Just Another Cognitive Function?

Chalmers insists that consciousness cannot be explained in functional terms. He claims that reducing consciousness (as we experience it) to a functional mechanism will *never* solve the hard problem:

> The easy problems are easy precisely because they concern the explanation of cognitive abilities and functions. To explain a cognitive function, we need only specify a mechanism that can perform the function. The methods of cognitive science are well-suited for this sort of explanation, and so are well-suited to the easy problems of consciousness. By contrast, the hard problem is hard precisely because it is not a problem about the performance of functions. The problem persists even when the performance of all the relevant functions is explained … What makes the hard problem hard and almost unique is that it goes *beyond* problems about the performance of functions. To see this, note that even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience … there may still remain a further unanswered question: *Why is the performance of these functions accompanied by experience?* A simple explanation of the functions leaves this question open … Why doesn't all this information-processing go on "in the dark," free of any inner feel? (Chalmers, 1995).

In the passage just quoted, Chalmers draws attention to the fact that consciousness is not just a cognitive function. It is easy to agree with him. All cognitive functions (such as memory) are not intrinsically conscious. There does not have to be "something it is like" to remember. It is well-established that learning and memory can exert their effects without any "inner feel"; and the same applies to perception. Hence the title of (Kihlstrom's, 1996) celebrated review article: "Perception without Awareness of What Is Perceived, Learning Without Awareness of What Is Learned." The only exception to the rule is precisely what needs to be explained: namely the *conscious part* of cognition—the part that is left over when the performance of all the relevant functions is explained.

Why is experience left unexplained, even when we have explained the performance of all the relevant cognitive functions in its vicinity? Some philosophers assert it is because "consciousness has a first person or subjective ontology and so cannot be reduced to anything that has third-person or objective ontology" (Searle, 1997, p. 212). The hard problem would be trivial if all it boils down to is the fact that your own personal experience, here and now, is not reducible to human experience in general. All one would need to do, then, to solve the problem, would be to take the experiences of lots of individuals, average them, find the common denominator, and explain *that* in functional terms. Psychologists do this sort of thing all the time. But Chalmers is not asking something so trivial. He writes:

> Why is it that when electromagnetic waveforms impinge on a retina and are discriminated and categorized by a visual system, this discrimination and categorization is experienced as a sensation of vivid red? We know that conscious experience *does* arise when these functions are performed, but the very fact that it arises is the central mystery. There is an *explanatory gap* (a term due to Levine, 1983) between the functions and experience, and we need an explanatory bridge to cross it. A mere account of the functions stays on one side of the gap, so the materials for the bridge must be found elsewhere (Chalmers, 1995).

Leaving aside his apparent conflation of two different kinds of explanatory gap (between experience and physiology on the one hand and experience and function on the other) it now becomes apparent why Chalmers believes that even the latter gap is unbridgeable. *He is focusing on the wrong function.* An explanation of experience will never be found in the function of vision—or memory, for that matter—or in any function that is not inherently experiential.

The function of experience cannot be inferred from perception and memory, but it *can* be inferred from feeling. There is not necessarily "something it is like" to perceive and to learn, but who ever heard of an unconscious feeling—a feeling that you cannot feel?[6] If we want to identify a mechanism that explains the phenomena of consciousness (in both its psychological and physiological aspects) we must focus on the function of feeling—the technical term for which is "affect." That is why it is easy to agree that consciousness is not just another cognitive function. Cognition has long been distinguished from affect, and for good reason[7].

This focus on affect is far from arbitrary.

IN THE BEGINNING WAS THE AFFECT

Is Consciousness a Cortical Function?

The massive effort in recent times to identify the NCC—*The Scientific Search for the Soul*, as Francis Crick (1994) memorably called it—used vision as its model example. This was justified by the fact that the details of visual processing are better understood than those for any other modality of consciousness.

Crick's strategy was that the NCC for vision should be generalizable to other forms of consciousness. His reasoning was simple: it must be possible to isolate something going on somewhere in the visual brain when you are seeing consciously which is absent when you are seeing unconsciously, and this is the NCC for vision. Closer study of this NCC (whatever it turns out to be: activation of a specific type of neuron, or a specific neural

[6] Freud always insisted that 'unconscious affect' is an oxymoron: "It is surely of the essence of an emotion that we should be aware of it, i.e., that it should become known to consciousness. Thus, the possibility of the attribute of unconsciousness would be completely excluded as far as emotions, feelings, and affects are concerned" (Freud, 1915a, p. 177). He explains: "The whole difference arises from the fact that ideas are cathexes—basically of memory-traces—whilst affects and emotions correspond to processes of discharge, the final manifestations of which are perceived as feelings. In the present state of our knowledge of affects and emotions we cannot express this difference more clearly" (ibid., p. 178).

[7] Strachey called Freud's distinction between 'quotas of affect' and 'memory-traces of ideas' the "most fundamental of all his hypotheses." (Strachey (1962), p. 63)

network, or a specific frequency band, etc.) should eventually reveal how and why visual consciousness arises.

In Chalmers's opinion, Crick's strategy is only capable of solving the easy (correlational) part of the mind/body problem; it cannot solve the hard (causal) part. There are at least three further problems with Crick's strategy.

The first is that there cannot be any objects of consciousness without a *subject* of consciousness. You cannot experience objects (visually or otherwise) unless *you* are there to experience them. This calls into question whether the essence of conscious experience resides in any perceptual modality. What if the NCC resides in the thing which binds the objects of conscious perception—in the perceiver rather than the perceptions?

This problem need not be fatal for Crick, if it turns out that experiencing arises from some aggregate of, or some interaction between, etc., the various types of perception—as some theorists claim it does. The experiencing subject need not take the form of a homunculus; it might be distributed over the cortex and emerge through a mechanism akin to trans-cortical "association." That is how the nineteenth century German anatomists saw it, when they first formulated the cortico-centric conception of consciousness on the model of seventeenth and eighteenth century British empiricist philosophies of mind (see Meynert, 1884).

This leads to a second problem with Crick's strategy. When Munk (1878, 1881) identified occipital cortex as the locus of the mental aspect of vision (which, importantly, he—like Meynert and the British empiricists—equated with the capacity to form visual "memory images" or "ideas," as opposed to mere sensations) it seemed reasonable enough to generalize the principle—the principle that the cortex is the organ of the "mind" *so defined*—to the other modalities of perception[8]. The ensuing experimental findings confirmed the validity of this generalization (e.g., ablation of auditory cortex [in dogs, Munk's model species] produced "mind deafness," just as occipital lesions caused "mind blindness"—which was subsequently also confirmed in human clinical cases; see Solms et al., 1996).

If we equate mind with memory images (and the associations between them) then it comes as no surprise to learn that, when Munk's contemporaries ablated the whole cortex, the animals did not fall into *coma*; instead, they became *amnestic* (see Meynert, 1884, Chapter 3, for review). Subsequent studies have confirmed this observation in numerous animal species (e.g., Huston and Borbely, 1974). Consciousness persists in the absence of cerebral cortex, as does volitional behavior. As Damasio and Carvalho (2013, p. 147) put it: "Decorticated mammals exhibit a remarkable persistence of coherent, goal-oriented behavior that is consistent with feelings and consciousness".

The same facts are observed in congenitally decorticate (hydranencephalic) human beings. In view of the importance of this for our topic, I will cite a lengthy description:

> In the setting of the home environment upon which these medically fragile children are crucially dependent, they give proof of being not only awake, but of the kind of responsiveness to their surroundings that qualifies as conscious by the criteria of ordinary neurological examination (Shewmon et al., 1999). The report by Shewmon and colleagues is the only published account based upon an assessment of the capacities of children with hydranencephaly under near optimal conditions, and the authors found that each of the four children they assessed was conscious. [...] To supplement the limited information available in the medical literature on the behavior of children with hydranencephaly, I joined a worldwide internet self-help group formed by parents and primary caregivers of such children. Since February of 2003 I have read more than 26,000 e-mail messages passing between group members. Of these I have saved some 1,200 messages containing informative observations or revealing incidents involving the children. In October 2004 I joined five of these families for 1 week as part of a social get-together featuring extended visits to DisneyWorld with the children, who ranged in age from 10 months to 5 years. I followed and observed their behavior in the course of the many private and public events of that week, and documented it with 4 h of video recordings. My impression from this first-hand exposure to children with hydranencephaly confirms the account given by Shewmon and colleagues. These children are not only awake and often alert, but show responsiveness to their surroundings in the form of emotional or orienting reactions to environmental events [...] They express pleasure by smiling and laughter, and aversion by "fussing," arching of the back and crying (in many gradations), their faces being animated by these emotional states. A familiar adult can employ this responsiveness to build up play sequences predictably progressing from smiling, through giggling, to laughter and great excitement on the part of the child. The children respond differentially to the voice and initiatives of familiars, and show preferences for certain situations and stimuli over others, such as a specific familiar toy, tune, or video program, and apparently can even come to expect their regular presence in the course of recurrent daily routines. Though behavior varies from child to child and over time in all these respects, some of these children may even take behavioral initiatives within the severe limitations of their motor disabilities, in the form of instrumental behaviors such as making noise by kicking trinkets hanging in a special frame constructed for the purpose ("little room"), or activating favorite toys by switches, presumably based upon associative learning of the connection between actions and their effects. Such behaviors are accompanied by situationally appropriate signs of pleasure or excitement on the part of the child, indicating that they involve coherent interaction between environmental stimuli, motivational-emotional mechanisms, and bodily actions [...] The children are, moreover, subject to the seizures of absence epilepsy. Parents recognize these lapses of accessibility in their children, commenting on them in terms such as "she is off talking with the angels," and parents have no trouble recognizing when their child "is back." [...] The fact that these

[8]At that time, "mind" and "consciousness" were synonymous. Despite his many disagreements with Meynert, Freud endorsed the view that consciousness is nothing more than "a sense organ for the perception of psychical qualities" (1900, p.615) and, moreover, that this "sense organ" was located in the *cerebral cortex*: "We have merely adopted the views on localization held by cerebral anatomy, which locates the 'seat' of consciousness in the cerebral cortex—the outermost, enveloping layer of the central organ. Cerebral anatomy has no need to consider why, speaking anatomically, consciousness should be lodged on the surface of the brain instead of being safely housed somewhere in its inmost interior." (Freud, 1920, p. 24). This cortical localization applied even to the affective aspect of consciousness (see Freud, 1940, pp, 161-2).

> children exhibit such episodes would seem to be a weighty piece of evidence regarding their conscious status (Merker, 2007, p. 79).

"Associative learning of the connection between actions and their effects" does not imply the experience of "memory images," but one must surely conclude that experience itself is not a cortical function. The ABCs of behavioral neuroscience demand that if a function (or critical component function) is localized in a particular structure, then ablation of that structure must result in loss of that function. In the case of consciousness in relation to the cerebral cortex, this critical test is failed.

I am aware that some readers will wonder about the above usage of the term "consciousness" (i.e., in what sense are these animals and children "conscious"); and they might invoke the epistemological problem of other minds (how do we *know* they are conscious). Before addressing these questions, let us consider a third problem with the cortico-centric approach.

The third problem is that there is a brain structure which *does* pass the critical test just mentioned. This structure is located not in the cortex but the brainstem.

The seminal observations were made in cats by Moruzzi and Magoun (1949), and confirmed in humans by Penfield and Jasper (1954). Consciousness is obliterated by focal lesions of the brainstem core[9]—in a region conventionally described as the extended reticulothalamic activating system (ERTAS). Recent findings indicate that the smallest lesions within the brainstem which cause total loss of consciousness (i.e., coma) are located in or near the parabrachial nuclei of the pons (Parvizi and Damasio, 2003; Golaszewski, 2016).

Why, then, did Crick and his followers not look for the NCC in the brainstem? The answer is: for reasons of convention. After Moruzzi & Magoun failed to confirm a major prediction arising from the classical theory, namely that deprivation of sensory inputs to cortex should result in loss of consciousness (e.g., sleep)[10], they did not abandon the theory; instead they introduced a distinction between the "contents" and "level" of consciousness. This saved the old theory. The contents (the *qualia* of consciousness) were thereby still assigned to the cortex, and a new level-regulating function (the *quantity* of arousal or wakefulness, measured on a 15-point scale) was assigned to the ERTAS.

This assignment continues to this day. Crick's closest collaborator, Christof Koch, therefore says of the deep brainstem nuclei that "they are *enablers* [of consciousness] but not content-providers" (Koch, 2004, p. 93, emphasis added). This takes us back to the question asked above: in what sense are decorticate animals and children conscious? Do they display *blank* wakefulness, devoid of content and quality, or is there "something it is like" to be them?

Does it Feel Like Something to be Awake?

The conclusion of the argument being set out here may be stated in advance: the so-called level of consciousness is a function of variational *free energy*. Free energy in thermodynamic terms entails *entropy*, which in information-theoretic terms is *surprisal* (and *uncertainty*), which in neurophysiological terms is *arousal* (see Solms and Friston, 2018) Arousal underpins *wakefulness*. Later, these equivalencies will enable us to approach the Helmholtzian ideal of describing "the specific way or form of the action [of consciousness] by means of the physical-mathematical method."[11] As Pfaff (2006) says: "Because CNS arousal depends on surprise and unpredictability, its appropriate quantification depends on the mathematics of information" (p. 13).

The question at hand concerns the nature of the "consciousness" displayed by decorticate animals and children. Consistent with what Damasio and Carvalho (2013) said about animals, and with Shewmon, Holmse and Byrne's (1999) findings, Merker (2007) observed that hydranencephalic children show "emotional or orienting reactions to environmental events." Moreover, "they express pleasure by smiling and laughter, and aversion by 'fussing,' arching of the back and crying (in many gradations), their faces being animated by these emotional states." The states include "smiling, through giggling, to laughter and great excitement on the part of the child." These children also "show preferences for certain situations and stimuli over others." And their "behaviors are accompanied by situationally appropriate signs of pleasure or excitement."

One surely must conclude that it does feel like something to be these children. By any reasonable standard[12], one would have to accept that they—like decorticate animals—show *basic emotions*. In fact, decorticate animals display excessive emotionality (Huston and Borbely, 1974), as do human patients who suffer damage to the cortical structures that exert inhibitory control over the ERTAS and limbic system (Harlow, 1868).

These observations may be linked with the fact that deep brain stimulation (DBS) of centrencephalic structures, such as the ERTAS and periaqueductal gray (PAG), and of the limbic circuits arising from them, generates powerful affective responses (see Panksepp, 1998, for detailed review). Importantly, in relation to the question concerning how we know these patients are conscious: in DBS of human beings, they declare these subjective states in words (e.g., Blomstedt et al., 2007). Within the confines of the epistemological problem of other minds (whereby one

[9]Ironically, in light of Freud's comment cited above, consciousness is *not* located "on the surface of the brain [but is instead] safely housed somewhere in its inmost interior." (Freud, 1920, p. 24)

[10]Cf. Meynert (1884) assertion: "The motor effects of our consciousness reacting upon the outer world are not the result of forces innate in the brain. The brain, like a fixed star, does not radiate its own heat: it obtains the energy underlying all cerebral phenomena from the world beyond it" (English trans., p. 160). Freud's views on this important point vacillated (see Solms and Saling, 1990).

[11]It will likewise enable us to approach the young Freud (1950 [1895]) unrequited aspiration "to represent psychical processes as quantitatively determinate states". Cf. his earlier remark to the effect that quotas of affect "possess all the attributes of a quantity (*though we have no means of measuring it*), which is capable of increase, diminution, displacement and discharge" (Freud, 1894, p. 60, emphasis added).

[12]The reasonable criterion here must be the same as it is for any other scientific question, namely, are predictions from the hypothesis (that these animals and children are conscious) *disconfirmed* or not? see (Panksepp et al., 2016).

can never know for certain whether anyone other than oneself is conscious) there can be no higher standard of proof for the inference that upper brainstem and limbic circuits generate affects[13]

This conclusion is further supported by the fact that drugs acting on the neuromodulators sourced in the ERTAS nuclei (serotonin, dopamine, noradrenaline, acetylcholine) have powerful effects on mood and anxiety, etc.—which is why they represent the mainstay of psychopharmacology today (Meyer and Quenzer, 2005). In other words, most psychotropic medications act via the ERTAS.

It is legitimate to say that affects are *generated* in these subcortical structures for the reason that the same effects can be observed in the absence of cortex. This contradicts the prevailing view that these nuclei merely "enable" the cortex to feel (Koch, 2004). It is noteworthy in this regard that patients with total destruction of the very structures which are specifically identified by cortico-centric theorists of affect—namely the prefrontal lobes and insula (e.g., Craig, 2009; LeDoux and Brown, 2017)—not only report preserved feeling states, but, as mentioned already, they display excessive emotionality (see Damasio et al., 2012)[14].

Although many cognitive scientists still must be weaned of the view that the cerebral cortex is the seat of consciousness (see Panksepp et al., 2016, for a lengthy discussion of this controversy), the weight of evidence for the alternative view that the arousal processes generated in the upper brainstem and limbic system feel like something in and of themselves, is now overwhelming. Coupled with the huge body of evidence suggesting that cortical (cognitive) functions are *not* intrinsically conscious (see Bargh and Chartrand, 1999, for review) one is led to the conclusion that the classical German anatomists were right: the cortex is merely a repository of "memory images." Cortex evidently provides "random-access memory" space (Solms and Panksepp, 2012, Ellis and Solms, 2018). This conclusion is consistent with the radical plasticity of cortex; so much so that the right hemisphere can take over the functions of the left, entirely, if it is removed early enough (Pulsifer et al., 2004); and when the optic nerve is redirected to auditory cortex, it learns to see (Sharma et al., 2000).

This line of thinking will be extended in section 'Consciousness Arises Instead of a Memory-Trace,' where it is argued that cortex *stabilizes* consciousness rather than generates it; i.e., that cortical functioning binds affective arousal, and thereby transforms it into conscious cognition.

It is undeniable that a hierarchical dependency relation exists between the cortical type of consciousness and the upper brainstem type. This is not a controversial claim; it is precisely what is meant by the conventional assertion that the ERTAS "enables" consciousness. In the absence of brainstem arousal there cannot be cortical consciousness, but the converse does not apply. Since these simple facts meet the gold standard for parsing neuropsychological functions—namely the principle of "double dissociation" (Teuber, 1955; see footnote 13)—we must conclude that consciousness is generated in the upper brainstem.

If core brainstem consciousness is the primary type, then consciousness is fundamentally *affective* (see Panksepp, 1998; Solms, 2013; Damasio, 2018). The arousal processes that produce what is conventionally called "wakefulness" constitute the experiencing subject. In other words, *the experiencing subject is constituted by affect.*

This reformulation of elemental consciousness has major ramifications for its functional mechanism, underscoring the conclusions reached at the end of section 'The Problem With The Hard Problem'. It is perfectly reasonable to ask why visual information-processing doesn't go on in the dark, without any inner feel, but it is perverse to ask why affective arousal doesn't do so. How can affective arousal (i.e., the arousal of feeling) go on without any inner feel?

Why Do We Feel?

Current theoretical efforts to answer this question were initiated by Damasio (1994), who identified feeling with registering states of the body—within a biological scale of values—whereby pleasurable vs. unpleasurable feelings register improving vs. deteriorating chances of survival and reproductive success[15]. On Damasio's theory, that is why we feel. His theory was substantially enhanced when he incorporated (Panksepp, 1998) findings to the effect that feelings are generated not in the cortex but the brainstem (and limbic system; see Damasio, 2010) and that the circuits in question do not register only here-and-now states (or "as if" states; Damasio, 1994) of the autonomic and sensory body, but also intrinsic brain states: brain systems for instincts[16] like attachment, rage and play (see Damasio, 2018). The shift downward to the brainstem enabled Damasio (like Panksepp before him) to recognize that *the elemental form of consciousness is an extremely primitive function*. My own contribution to these theoretical efforts came relatively late in the day (Solms and Panksepp, 2012) and they revolved mainly around the precise relationship between

[13]Of course, this does not imply that other structures do not also participate, even (and importantly) including some beyond the confines of the nervous system. The function of affect is being 'localized' in the conventional sense demanded by Teuber's "double-dissociation" paradigm, which states that if function A is lost with damage to structure X but not structure Y, and function B is lost with damage to structure Y but not structure X, then functions A and B are two independent functions. (Here A = consciousness; B = cognition; X = brainstem; Y = cortex).

[14]Freud shared the cortico-centric view that even affects are felt only when the underlying 'psychical energies' arouse what he termed the 'inner surface' of the system Pcpt.-Cs. in the cerebral cortex (see footnote 8 above)

[15]This view was not original; it coincided almost exactly with Freud's view to the effect that "oscillations in the tension of instinctual needs [. . .] become conscious as feelings in the pleasure-unpleasure series" (1940, p. 198). Damasio (1999) acknowledged Freud's priority.

[16]There is no generally-agreed-upon definition of 'instinct' but it should be noted that the term is being used here in the mainstream biological sense rather than the Freudian one (which, incidentally, arose from a mistranslation of the German term Trieb; see Solms, 2018c).

homeostasis[17] and feeling (Solms, 2013, 2018a; Solms and Friston, 2018).

To be clear: I do not claim (and nor did Panksepp or Damasio)[18] that feeling arises from homeostasis in and of itself. I do not believe that thermostats are conscious. I do not even claim that all living creatures are conscious (although all living creatures are homeostatic). Even in human beings, homeostatic mechanisms which are totally devoid of consciousness are operative. The regulation of blood pressure is a clinically notorious example. In fact, one may go much further: like Freud (and just about everyone else these days) I do not claim that all human *mental* functions are conscious. This has important implications for philosophers like Nagel and Chalmers, who sometimes forget that "subjectivity" and "consciousness" are not synonymous words (This fact is especially problematical for Chalmers's panpsychism; see Chalmers, 1995, 1996).

What I am claiming is something else: feeling enables complex organisms to register—and thereby to regulate and prioritize through thinking and voluntary action—deviations from homeostatic settling points *in unpredicted contexts*. This adaptation, in turn, underwrites learning from experience. In predictable situations, organisms may rely on automatized reflexive responses (in which case, the biologically viable predictions are made through natural selection and embodied in the phenotype; see Clark, 2016). But if the organism is going to make plausible *choices* in novel contexts (cf. "free will") it must do so via some type of here-and-now assessment of the relative *value* attaching to the alternatives (see Solms, 2014).

Crucially, in this process, the organism must stay "ahead of the wave" of the biological consequences of its choices (to use the analogy that gave Andy Clark's (2016) book its wonderful title: *Surfing Uncertainty*):

> To deal rapidly and fluently with an uncertain and noisy world, brains like ours have become masters of prediction—surfing the waves of noisy and ambiguous sensory stimulation by, in effect, trying to stay just ahead of the place where the wave is breaking (p. xiv).

The proposal on offer here is that this imperative *predictive* function—which bestows the adaptive advantage of enabling organisms to survive in novel environments—is performed by feeling (see section 'To Be Precise' below for clarification of the pivotal role of *context* in the prioritization of affects, and thereby the "flavoring" of consciousness). On the present proposal, this is the causal contribution of qualia (see Solms and Friston, 2018).

Affective qualia are accordingly claimed to work like this: deviation away from a homeostatic settling point (increasing uncertainty) is felt as unpleasure, and returning toward it (decreasing uncertainty) is felt as pleasure[19]. There are many types (or "flavors") of pleasure and unpleasure in the brain (Panksepp, 1998)[20]. The type identifies the need at issue, which enables the organism to minimize computational complexity (i.e., to focus on the matter at hand—rather than its organismic state as a whole—and thereby to minimize metabolic expenditure; see Solms and Friston, 2018, footnote 7). All needs cannot be felt at once. The prioritization of needs—i.e., the determination as to which need will be felt—must obviously depend crucially upon *context* (i.e., needs in relation to other needs, and needs in relation to opportunities)[21]. Feeling is therefore extended onto exteroception (i.e., it is contextualized: "I feel like this about *that*") and transformed into cognitive consciousness (i.e., it is "bound"; see section 'Consciousness Arises Instead of a Memory-Trace'). This in turn gives rise to voluntary action—and what we loosely call *thinking*—and, over longer time-scales, to learning from experience (Thinking, as Freud taught us, entails virtual action rather than real action, and thereby saves lives)[22].

Consciousness (thus defined) is a biological imperative; it is the vehicle whereby complex organisms monitor and maintain their functional and structural integrity in unknown situations. The inherently subjective and qualitative nature of this auto-assessment process explains "how and why" it [consciousness] feels like something to the organism, for the organism (cf. Nagel, 1974). Specifically, increasing uncertainty in relation to any biological imperative *just is* "bad" from the (first-person) perspective of such an organism—indeed it is an existential crisis—while decreasing uncertainty *just is* "good." This provides a very important clue as to how the "hard problem" may be solved. Consciousness adaptively determines which uncertainties must be felt (i.e., prioritized) in any given context. In short, consciousness is *felt uncertainty*. We will see shortly how and why the first person perspective arises.

[17] Many commentators forget that the term "homeostasis" was only introduced into biology in 1926. Freud conceptualized the same function as "drive." In this respect, the following extract from Solms (2013, pp. 79-80) serves as a summary of the present article: "I define drive as 'a measure of the demand made upon the mind for work in consequence of its connection with the body" (Freud, 1915a, p. 122), where the "measure" is the degree of deviation from a homeostatic set-point (with implications for survival and reproductive success). I do not believe that this deviation itself is something mental, but the 'demand' it generates is felt in the pleasure-unpleasure series. This (felt demand) is affect, which in my view is the origin of mind. The "work" that flows from affect is cognition, the functional purpose of which is to reduce affect—that is, to reduce prediction error (free energy). The purpose of cognition is to bring the world into line with our predictions and our predictions into line with the world. This centrally involves learning."

[18] Damasio (2018) attributes feeling states only to creatures with nervous systems (see Solms, 2018b).

[19] In the view on offer here, therefore, unlike Freud's, the drive *is* the feeling (Solms, 2013). Drive literally brings the mind into being. Before the drive is felt it is not a drive – it is simple homeostasis, which can be regulated by autonomic reflexes and behavioral stereotypes. The present view also differs from Freud's in conceptualizing pleasure-unpleasure as deviations to and from a settling point, as opposed to Freud's continuum, and in conceptualizing Nirvana as that settling point rather than something 'beyond' the pleasure principle (see Solms, 2018b).

[20] The conflicting demands of the different needs that these many "flavors" represent underpins mental conflict, and (equally importantly) accounts for the many behaviors one sees in nature—and in psychopathology—which are by no means obviously "self-preservative."

[21] Panksepp (1998) and Merker (2007) provide cogent evidence for the view that this prioritization process pivots around a midbrain "decision triangle" (Panksepp calls it the "SELF") whereby *needs* are registered in the periaqueductal gray (PAG) and *opportunities* in the superior colliculi.

[22] Cf. Freud's notion that thinking is interposed between drive and action. The contents of this paragraph are necessarily overly dense. These highly complex issues require more space than a journal article allows. See Solms, in press, for a more detailed explication.

At this point, however, we must confront what philosophers term the "conceivability problem."

The function I have just described could conceivably be performed by non-conscious "feelings" (cf. philosophical zombies)—if evolution had found another way for living creatures to pre-emptively register and prioritize (to themselves and for themselves) such inherently qualitative existential dynamics in uncertain contexts. But the fact that something can conceivably be done differently doesn't mean that it is not done in the way that it is in the vertebrate nervous system. In this respect, consciousness is no different from any other biological function. Ambulation, for example, does not *necessarily* require legs (As Jean-Martin Charcot said: "Theory is good, but it doesn't prevent things from existing'; Freud, 1893, p. 13). It seems the conceivability argument only arose in the first place because we were looking for the NCC in the wrong place. One suspects the problem would never have arisen if we had started by asking how and why feelings (like hunger) arise in relation to the exigencies of life, instead of why experience attaches to cognition.

In the next section, I will reduce the function of consciousness to its formal essence. But I want to conclude the present section with a brief description of its anatomical realization:

Body-monitoring nuclei in the spinal cord (dorsal root ganglia), upper brainstem and diencephalon (e.g., solitary nucleus, area postrema, parabrachial nucleus, circumventricular organs, and hypothalamus) can only go so far in terms of meeting endogenous needs through internal (autonomic) adjustments. Beyond that limit, *external* action is called for. At that point, autonomic reflexes become *drives*. That is, interoceptive (mainly medial hypothalamic) "need detectors" trigger not only autonomic reflexes but also—following the crucial prioritization process performed by the midbrain "decision triangle" (see footnote 21 above)—feelings of hunger, thirst, etc. Through a final common pathway of ERTAS arousal these drives typically[23] trigger dopaminergically-mediated "foraging" behaviors (viz., the behaviors that Panksepp (1998) calls "SEEKING" and Berridge (1996) calls "wanting"). Foraging reflects a phylogenetically determined prediction, namely the prediction that whatever I need will be found out there in the world. The difference between Panksepp's "SEEKING" (i.e., objectless drive) and Berridge's "wanting" (i.e., goal-oriented motivation) reflects the influence of learning upon the primary instinctual mechanism of desire—whereby affective SEEKING becomes cognitive "wanting" (through need/satisfaction matching)[24]. This facilitates the formation of LTM cause/effect relations between particular needs and their adequate aims and objects, which in turn yields the iterative "reward prediction error" cycle that codes ongoing learning from experience (see Schultz, 2016).

Fortunately, living organisms are not required to learn everything about the world from scratch. Each phenotype is endowed with innate predictions concerning biologically significant situations it is certain to encounter[25]. Panksepp (1998) terms these "emotional" and "sensory" affects (but it is important to recognize that the word "affect" is only justified to the extent that the relevant instinctual and reflexive predictions are felt, i.e., to the extent that they yield residual uncertainties, which require choice and learning from experience). Examples of "emotional" affects (each of which is marked by its own command neuromodulators and receptor types) are fear, rage, attachment and play; and examples of such "sensory" affects are pain, surprise and disgust (see Panksepp, 1998). Fear behaviors (freezing and fleeing), for example, are innate predictions; but each individual has to learn *what* to fear and *what else* might be done in response. What vertebrates do to meet their needs always consists in a combination of innate and learned behaviors.

The residual uncertainty (unmet needs—i.e., unsolved problems—of various types) arising from each such cycle of behavior is auto-evaluated, in the manner described above, by mechanisms located mainly in the PAG—the terminus of all affective circuitry[26]. Merker (2007) accordingly describes the PAG as part of a "synencephalic bottleneck," where perception, action and affect come together, and choices are made as to "what to do next[27]." (It is important to recognize that the terminal location of the PAG in the cycle just described renders it *functionally* "supra-cortical," notwithstanding the fact that it is *anatomically* sub-cortical; see Merker, 2007). PAG activity, then, results in revised perception/action selection, via ERTAS (and more specific higher limbic) neuromodulatory adjustments. This is how simple feeling becomes "feeling *about that*[28]."

Note that the evaluation cycle just described entails ongoing assessment of environmental events *and* the internal milieu (via body monitoring nuclei)—both of which are "external" to the nervous system—although, for obvious biological reasons, internal uncertainties will almost always trump external ones (Imagine the consequences of back-ranking changes in oxygenation or hydration or thermoregulation). That is why consciousness is quintessentially affective.

[23]I say 'typically' because foraging is commonly the most adaptive response to contextual uncertainty. However, all manner of other instincts may be selected, which are so conditioned through learning from experience, that they are frequently no longer recognizable as instincts at all. (Cf. what is said below about learning in relation to the SEEKING instinct, which serves as a model example.)

[24]This seems to be identical with Freud (1950 [1895]) conception of the cognitive effects of 'experiences of satisfaction'; i.e., wishful cathexis, etc. For the role played by opioids in such experiences, see Berridge et al. (2009). Panksepp (1998) and Schultz (2016) offer distinctly different accounts of the role played by dopamine.

[25]Freud endorsed the concept of basic emotions, although he classified them differently from how we do today, and he conflated them with his conception of primal phantasy—which entailed the untenable notion of inherited episodic memories. See Freud, 1916-17, p. 395.

[26]Focal lesions of the PAG produce persistent vegetative states, and DBS there elicits powerful affects of various kinds—not only negative ones—depending upon which part of the PAG is stimulated.

[27]See footnote 21 above. Freud (1900), too, placed the system Cs. at the motor end of the apparatus, but he evidently had *cortical* motor mechanisms in mind.

[28]Freud (1900), too, pictured a functional overlap between Cs. interoception and Pcpt. exteroception, and eventually he combined the two systems under the single rubric "Pcpt.-Cs." (Freud, 1917). However, once again, he clearly had *cortical* systems in mind.

TO BE PRECISE

How Does Homeostasis Arise?

If consciousness arises through a homeostatic mechanism, as the above physiological[29] considerations suggest, then a lot rides on the question: how does homeostasis arise? The answer to this question should lead to the abstraction we are looking for (i.e., the abstraction that transcends psychological and physiological "appearances").

According to Friston (2013) the answer is *free-energy minimization*. For self-organizing systems—including all living things, like us—to exist, they must *resist entropy* (quantified as free energy, but see below for the important role of precision weighting)[30]. That is, self-organizing systems can only persist over time by occupying "preferred" states—as opposed to being dispersed over all possible states, and thereby dissipating. This is a fundamental precondition of life—and indeed any self-organization. We need not concern ourselves here with how life arises. However, grounding the mechanism of consciousness in the essential prerequisites for life is not a bad starting point, since it is generally assumed that all conscious things are alive—although not all living things are conscious.

For a system to resist entropy, three conditions must be met: (i) There must be a boundary which separates the internal and external states of the system, and thereby insulates the system from the world. Let's call the former states "the system" and the latter states "the not-system"—rather than "the world," for reasons that will soon be explained. (ii) There must be a mechanism which registers the influence of dissipative external forces—i.e. the free energy. Let's call this mechanism the "sensory states" of the system. (iii) There must be a mechanism which counteracts these dissipative forces—i.e. which binds the free energy. Let's call this mechanism the "active states" of the system, such as motor and autonomic reflexes[31].

According to Friston (2013), these functional conditions—which enable self-organizing systems to exist and persist over time—emerge naturally (indeed necessarily) within any ergodic[32] random dynamical system that possesses a *Markov blanket*[33]. This blanket establishes the boundary conditions above and is a probabilistic construct that depends upon what influences what (and what doesn't influence what). The Markov rules of causal influence provide the prerequisite (i) separation between the system and the not-system (i.e., the blanket itself), and equip the former with (ii) receptor capacities (the sensory states of the blanket) and (iii) effector capacities (the active states of the blanket). It is important to recognize that these sensory and active capacities are properties of the blanket—not of the states they interact with—which implies that the system insulated by a Markov blanket can only "know" states of the not-system *vicariously*. In other words, external states can only be "inferred" by the system—on the basis of "sensory impressions" upon the Markov blanket.

In fact, it is *essential* for external states to be inferred by the system if dissipative forces are to be resisted. This implies that the system must incorporate a *model* of the world, which then becomes the basis upon which it acts. Such models—like all models—are imperfect things. They can (and must) be improved in the light of unfolding evidence. In other words, the inferences the model generates for the system about conditions outside (inferences formed on the basis of the sensory consequences of its actions) take the form of predictions, and these predictions must be constantly tested and revised[34]. Thus, perception and action entail ongoing processes of *hypothesis testing*, whereby the system updates its model—its "beliefs[35]"—over time. This imperative of negentropic self-organizing systems is, in a nutshell, what Friston calls "active inference." Mathematically, the quality of this model corresponds to model evidence; namely the probability of sensory fluctuations under the model. In this setting, free energy provides a function of sensory states that must decrease when model evidence increases. In other words, self-organization—and implicitly any form of homoeostasis—can be cast as minimizing free energy (or, more simply, self-evidencing).

One must add that if the self-organizing system at issue is a nervous system, then—odd as this may sound—it is important to recognize that all other bodily systems (e.g., the viscera) are "external" to the nervous system[36]. Nervous systems sense, represent and act upon all other bodily systems (both vegetative

[29] I have emphasized the physiological considerations over the psychological ones in this account. The parallel commentary in these footnotes draws attention to the fact that the physiological inferences we have reached strongly resemble the psychological inferences that Freud was led to. For him, feelings (the pleasure principle) were the bedrock of mental life—including cognition.

[30] Freud (1920) encapsulated this fundamental biophysical dynamic in his second drive theory. Before that, he formulated it as a compromise—the "constancy principle"—which he imagined as being effected by a reticulum of "constantly cathected neurons" (Freud, 1950 [1895]), the "great reservoir" of his later "ego," the "bound energy" of which gave negentropic power to the "secondary process." By this I mean the capacity to inhibit neuronal discharge (called "freely mobile energy" in Freud's terminology), which he equated with the action of the Second Law (see his principle of "neuronal intertia," the direct ancestor of the "death drive"). In Friston's predictive processing framework, this same negentropic power is attributed to predictive neuronal assemblies (which are directly equivalent to Freud's LTM Ψ neurons) which inhibit transmission of sensory signals—Freud's STM Φ neurons—thereby minimizing "prediction error" and all the entropic perturbations it gives rise to, measured as "free energy". Cf. (Carhart-Harris and Friston, 2010).

[31] Cf. Freud's concepts: (i) "Q screens" or "stimulus barriers", (ii) "ϕ neurons" or "system Pcpt" and (iii) "M neurons" or "system Cs," respectively. Incidentally, most Freud scholars do not seem to realize that Q, in thermodynamics, quantifies heat.

[32] "Ergodicity" is a statistical property, whereby the average of any measurable function of a random dynamical system *converges* over a sufficient period of time. In short, dynamical systems that possess measurable characteristics over periods of time must be (nearly) ergodic.

[33] A "Markov blanket" induces a statistical partitioning of internal and external states, and *hides* the latter from the former. The Markov blanket itself consists in two sets ("sensory" and "active" states) which influence each other in a circular fashion: external states cause sensory states which influence—but are not influenced by—internal states, while internal states cause active states which influence—but are not influenced by—external states.

[34] Freud would have called such predictions "unconscious phantasies."

[35] This sensory sampling process is reminiscent of Freud's image of the system Ucs periodically palpating the system Pcpt-Cs with cathectic feelers (Freud, 1925b, p. 231). 'Beliefs', in the sense used here, are taken to be probability distributions whose parameters or sufficient statistics correspond to system states.

[36] Freud (1950 [1895]) speaks of "the somatic element itself" generating Q (which he designates Qη) by virtue of "an increasing complexity of the interior of the organism." (p. 297).

and sensory-motor ones) in just the manner I have described. Nervous systems co-evolved with the other systems due to increasing complexity of organisms, which (complexity) requires orchestration of the multiple homeostatic demands arising from the various systems. Nervous systems are therefore meta-systems, performing meta-homeostatic functions on behalf of the entire body. Homeostatic regulation of the organism as a whole is delegated, as it were, to the nervous system.

In summary, homeostasis is explained by the causal dynamics mandated by the very existence of Markov blankets; in terms of which self-organizing systems generate a type of work that binds free energy and maintains the system in its typically occupied ("preferred" or "valued") states. The concept of preferred states of self-organizing systems is identical with the concept of homeostatic settling points. The mathematical formulations quantifying the relevant dynamics of self-organizing systems need not be reproduced here (see Friston, 2013); since they concern the prerequisites of life in general rather than those for *consciousness* in particular. I will introduce the equations that are critical for our purposes in the next subsection.

Hopefully it is clear from the forgoing that although I have used quasi-physiological terms like "sensory" and "motor," and quasi-psychological ones like "knowing," "inference," "belief," "value" and "prediction," the actual mechanisms I have described are simultaneously physiological *and* psychological ones. This (their abstract ontology) is their primary virtue, in light of what I said in section 'The Problem With the Hard Problem'. As we shall now see, the very same abstractions can be extended to explain the function of consciousness in both its (psychological and physiological) manifestations. Indeed, that is why one is justified to use quasi-physiological and quasi-psychological terms for these mechanisms.

Now we come to the crux of the matter.

How Does Consciousness Arise?

I first expressed the view in 1997 that the problem of consciousness will only be solved if we reduce its psychological and physiological manifestations to a single underlying abstraction (Solms, 1997)[37]. It took me many years to realize that this abstraction revolves around the dynamics of free energy and uncertainty (Solms, 2013, 2014).

Free energy minimization is the basic function of homeostasis, a function that is performed by the same brainstem nuclei that I was led to infer—like others, on independent (clinico-anatomical) grounds—were centrally implicated in the generation of consciousness. In other words, the functions of homeostasis and consciousness are realized physiologically in the very same part of the brain. This insight led to the collaborative work that enabled Friston and me to expand the variational free energy formulation of the mechanism of homeostasis to explain the mainspring of consciousness itself (Solms and Friston, 2018)[38].

Readers may have noticed already that the dynamics of a Markov blanket generate two fundamental properties of minds—namely (elemental forms of) *selfhood* and *intentionality*. It is true that these dynamics also generate elemental properties of bodies—namely an *insulating membrane* (the ectoderm of complex organisms, from which the neural plate derives) and *adaptive behavior*. This is a remarkable fact. It underpins dual-aspect monism.

Section 'In the Beginning Was the Affect' focused mainly on the anatomy and physiology of homeostasis; now we are also clarifying its psychology, by explicating the deeper mechanism. Foundational to what we call psychology is the *subjective* observational perspective. The fact that self-organizing systems must monitor their own internal states in order to persist (that is, to exist, to survive) is precisely what brings *active* forms of subjectivity about. The very notion of selfhood is justified by this existential imperative. It is the origin and purpose of mind.

Selfhood is impossible unless a self-organizing system monitors its internal state in relation to not-self dissipative forces. The self can only exist in contradistinction to the not-self. This ultimately gives rise to the philosophical problem of other minds. In fact, the properties of a Markov blanket *explain* the problem of other minds: the internal states of a self-organizing system can only ever register hidden external (not-system) states vicariously, via the sensory states of their own blanket.

We have seen that minds emerge in consequence of the existential imperative of self-organizing systems to monitor their own internal states in relation to potentially annihilatory, entropic forces[39]. Such monitoring is an inherently value-laden process. It is predicated upon the biological ethic (which underwrites the whole of evolution) to the effect that survival is "good." This imperative is formalized in terms of free-energy minimization.

Such negentropic dynamics of self-organizing systems are the absolute precondition for the evolution of minds. However, there is nothing about these dynamics which distinguishes conscious from unconscious mental processes. Put differently, there is nothing about such proto-mental dynamics which explains the emergence of feeling, as opposed to the exigencies of life. It is true that the dynamics described above revolve around value, but the values in question could—in principle—still be expressed in purely quantitative terms (e.g., $10 > 9$). There is no necessity to introduce qualitative terms into the dynamics of free energy minimization.

What is it then, that underwrites the transition from unconscious (quantitative, "proto-mental") states to conscious (qualitative, truly "mental") ones? It seems the transition revolves

[37] Freud's unifying abstraction was the "mental apparatus". The philosophical implications of his oft-repeated insistence that the instrument of the mind is unconscious "in itself" are not sufficiently appreciated (see Wakefield, 2018). Hence his laconic remark: "the unconscious is the proper mediator between the somatic and the mental, perhaps the long-sought 'missing link'" (letter to Georg Groddeck dated June 5, 1917; see Groddeck, 1977).

[38] When we did so, I experienced something similar to what Freud described more than a century before, when he wrote: "Everything seemed to fit together, the gears were in mesh, the thing gave one the impression that it was really a machine and would soon run of itself […] Of course, I cannot contain myself with delight." (Letter to Fliess of October 20, 1895; Freud, 1950 [1892-99]).

[39] Cf. Freud's formulation of narcissism ("hate, as a relation to objects, is older than love"; Freud (1915b), p. 139) which became the foundation of Melanie Klein's 'paranoid schizoid position'.

fundamentally around increasing complexity. This refers to complexity of a specific type, however, not just complexity of integrated information processing in general (cf. Tononi, 2012). On the self-evidencing view, complexity acquires a very specific meaning[40] (This follows from the fact that model evidence is the difference between accuracy and complexity. As model evidence is actively increased by minimizing free energy, the accuracy of predictions rises, with a concomitant increase in complexity. In other words, increasing model complexity is always licensed by an ability to make more accurate predictions).

Organisms evolve increasing self-complexity—for obvious adaptive reasons—as they diversify into (divide vegetative labor between) multiple sub-systems. For example, they evolve digestive vs. respiratory vs. thermoregulatory vs. immune systems. Each such specialized system is governed by a homeostatic imperative of its own. Metabolic energy balance, oxygenation, hydration, and thermoregulation (for example) are not the same things, although each of them contributes to the overall imperative of organism-wide free energy minimization. If the differential demands of the specialized homeostatic systems are going to be computed differentially (as they must) then it follows that increasing complexity requires some form of compartmentalization of quantities. Such compartmentalization can only be achieved through some form of *qualitative* differentiation between the sets of variables (e.g. 10 × X is worth more than 10 × Y; where X and Y are *categorical variables*). One can think of this compartmentalization as being something akin to a "color coding" or "flavoring" of the different data sets. This manifests in many different guises; from functional specialization in neuronal systems through to factorization of fundamental constructs that we use to model the world (e.g., "what" and "where" systems in the brain). As noted above, model evidence is the difference between accuracy and complexity, which requires increases in complexity to be nuanced (cf. Ockham's principle). Compartmentalization enables a simpler representation of what's going on "out there" in terms of external or non-self-states. Crucially, this sort of compartmentalization is essential for models that generalize to new situations.

In other words, the requirement for compartmentalization becomes a necessity when the relative value of the different quantities *changes* over time. For example: hunger trumps fatigue up to a certain value, whereafter fatigue trumps hunger; or hunger trumps fatigue in certain circumstances, but not others (i.e., 10 × X is currently [but not always] worth more than 10 × Y). Such changes require the system not only to compartmentalize its work efforts in relation to its different needs, but also to *prioritize* them over time.

This imperative reaches its nadir in the active states of the system, which inevitably produce a bottleneck. For example, organisms cannot eat and sleep simultaneously. Likewise, they cannot turn left and right at the same time. When it comes to action, executive choices must be made.

All these *contextual* factors become more prescient when one considers also how organisms survive in novel (unpredicted) environments. It is conceivable that an extremely complex set of algorithms could evolve (no matter how unwieldy they may become) to compute relative survival demands in all predictable situations, and to prioritize actions on this basis. But how does the organism choose between X and Y when the consequences of the choice are unpredictable? The physiological considerations discussed in the previous section suggest that it does so by *feeling* its way through the problem, where the direction of feeling (pleasure vs. unpleasure)—in the relevant modality—predicts the direction of expected uncertainty (decreasing vs. increasing)—within that modality[41].

In selecting the best course of action, we must call upon our model of the world to predict the consequences of some behavior in terms of the expected free energy. *Expected free energy just is uncertainty about the consequences of any putative action*. The imperative to minimize expected free energy therefore becomes necessary to choose actions that minimize uncertainty and realize familiar, preferred sensory states.

Before we consider what this might entail in formal, mathematical terms, I want to make clear that the evolutionary considerations we have just reviewed suggest a *graded* transition from proto-mental to mental states (i.e., from unconscious to conscious subjectivity). Subjective values (i.e., system-centric values) are computed at the level of autonomic homeostasis already. This implies a potential for hedonic valence. But the qualitatively felt aspect of hedonic value does not *have to* be registered by the self-organizing system until multiple such values must be differentially computed and prioritized in variable and novel contexts, where uncertainty itself becomes the primary determinant of action selection.

Computationally, such contextual factors are formalized in terms of precision-weighting. "Precision" is an extremely important aspect of active and perceptual inference; it is the *representation of uncertainty*. The precision attaching to a quantity estimates its reliability, or inverse variance (e.g., visual—relative to auditory—signals are afforded greater precision during daylight vs. night-time). Heuristically, precision can be regarded as the confidence afforded probabilistic beliefs about states of the not-system—or, more importantly, what actions "I should select."

This is the fundamental point made in Solms and Friston (2018). We were led to the conclusion that—whereas homeostasis requires nothing more than ongoing adjustment of the system's active states (M) and/or inferences about its sensory states (ϕ), in accordance with its predictive model (ψ) of the external world (Q) or vegetative body ($Q\eta$), which can be adjusted automatically on the basis of ongoing registrations of prediction error (e), quantified as free energy (F)—the contextual considerations just

[40] Technically, it is the relative entropy between posterior and prior beliefs or probability distributions over external or not-self states.

[41] A common source of confusion here is the fact that the dopaminergic SEEKING modality (discussed in Section 'In the Beginning Was the Affect') engages *positively* with uncertainty. Its innate non-declarative prediction translates as: 'engagement with a source of uncertainty provides maximal opportunities to resolve that uncertainty'. Therefore, in the case of this instinct, lack of engagement with uncertainty is "bad" (cf. anergia, abulia, anhedonia, hopelessness). The conceptual distinction in the affective neuroscience of our time between "appetitive" and "consummatory" pleasures removes the source of Freud's puzzlement in his lifelong attempts to establish a psychophysics of pleasure-unpleasure in relation to oscillations in the tension of drive needs.

reviewed require an additional capacity to adjust the precision weighting (ω) of all relevant quantities. This capacity provides a formal (mechanistic) account of voluntary behavior—of choice.

With the above quantities[42] in place, one can describe any self-organizing (i.e., self-evidencing) system with the following dynamics:

$$\frac{\partial}{\partial t}M = -\frac{\partial F}{\partial M} = -\frac{\partial F}{\partial e}\frac{\partial e}{\partial M} = \frac{\partial \Phi}{\partial M}\cdot\omega\cdot e \quad (1a)$$

$$\frac{\partial}{\partial t}Q = -\frac{\partial F}{\partial Q} = -\frac{\partial F}{\partial e}\frac{\partial e}{\partial Q} = -\frac{\partial \psi}{\partial Q}\cdot\omega\cdot e \quad (1b)$$

$$\frac{\partial}{\partial t}\omega = -\frac{\partial F}{\partial \omega} = \frac{1}{2}\cdot\left(\omega^{-1} - e\cdot e\right) \quad (1c)$$

Where free energy and prediction error are:

$$F = \frac{1}{2}\cdot\left(e\cdot\omega\cdot e - \log(\omega)\right) \quad (2)$$

$$e = \Phi(M) - \psi(Q) \quad (3)$$

A more detailed account of the thinking behind these broad-brushstroke equations can be found in Solms and Friston (2018) and in the background references contained therein.

Physiologically, precision is usually associated with the *postsynaptic gain* of cortical neurons reporting prediction errors. This is precisely the function of ERTAS modulatory neurons (see section In the Beginning Was the Affect). In this sense, precision can be associated—through free energy minimization—with *selective arousal* (and thus, as formalized by the three dependencies in equation 1, with action [1a], perception [1b], and affect [1c], respectively).

It is useful to appreciate that every prediction error neuron (or neuronal population) is equipped with a specific—and changing—postsynaptic gain, and thereby with an implicit representation of precision. Precision is not a single value; every sensation and action—and every hierarchical abstraction, including every prediction and ensuing error signal—must be equipped with a precision which has to be optimized.

From the above equations, it is also clear that precision (consciousness) *controls* the influence of prediction errors on action (motivation) and perception (attention). Conceptually, precision is a key determinant of free energy minimization and the enabling—or activation—of prediction errors. In other words, precision determines which prediction errors are selected and, ultimately, how we represent the world and our actions upon it.

In this sense, precision plays the role of Maxwell's daemon[43]–selecting the passage of molecules (i.e., sensory signals) to confound the Second Law of thermodynamics. In this analogy, consciousness is nothing more or less than the activity of Maxwell's daemon (i.e., the optimization of precision with respect to free energy). That is, in this analogy, consciousness does not correspond to the passage of molecules that are enabled by the daemon (i.e., the perceptual sequelae of message passing in cortical hierarchies) but rather to the activity of the daemon itself.

This distinction is what underlies the prejudice (of Koch and others) to the effect that neuromodulation merely "enables" conscious content. The conceptual breakthrough reported here revolves around the insight that the residual error in each action/perception cycle (registered in PAG, see section 'In the Beginning Was the Affect') is *felt* uncertainty—i.e., that each of the various categories (or flavors) of error possess affective "content" of their own. Here, unpleasure (within the modality at issue) means *increasing uncertainty* in the modality, and pleasure means that *things are turning out as expected*. This (felt uncertainty) causally determines the (ERTAS) adjustments of subsequent sensory-motor priorities and expectations (i.e., of ω). That is, it determines selective arousal. This is the heart of the matter.

Note that this proposal calls on the notion of *activating* expectations or representations in the sense that—in the absence of precision—prediction errors could fail to induce any neuronal response. In other words, without precision, prediction errors could be sequestered at the point of their formation in the sensory epithelia (or at whichever level in the predictive processing hierarchy they occur). Physiologically, these sorts of states are encountered every day; for example, in stereotyped behavioral automatisms and during sleep (Hobson, 2009; Hobson and Friston, 2014)[44].

The distinction between *interoceptive* and *exteroceptive* precision is central to this argument. If brains are sympathetic organs of inference, assimilating exteroceptive (sensory/motor) and interoceptive (vegetative) data through prediction, then their respective precision is about something (c.f. Brentano, 1874).

The proposal is that interoceptive precision is prioritized because the probabilistic beliefs attaching to what Panksepp calls homeostatic affects (e.g., hunger, thirst, sleepiness) cannot be overridden. Organismic beliefs at this level of the hierarchy are dictated by the phenotype, not by experience. This implies that everything which follows in the hierarchy, leading from the centrencephalic core to the sensorimotor periphery, is subordinated to affect. That is why I describe the adjustment of ω *per se* as "affect". Consciousness *itself* is affective. Everything else (from motivation and attention, leading to action and perception, and thereby to learning)—all of it—is a functional of affect. Affect *obliges* the organism to engage with the outside world,

[42] ω, precision; ψ, prediction; ϕ, perception; M, action; Q, {inferred] world; F, free energy; e, prediction error. Psychoanalytic readers will recognize some of these quantities from Freud, 1950 [1895]). We use the same symbols in recognition of the penetrating insights contained in his "Project," although it has become necessary—in line with some further insights recorded in the footnotes above—to use them slightly differently from what Freud had in mind.

[43] Maxwell's daemon is a thought experiment created by James Clerk Maxwell to suggest how the second law of thermodynamics might be violated: in brief, a daemon controls a small door between two chambers of gas. As gas molecules approach, the daemon opens and shuts the door, so that fast molecules pass to the other chamber, while slow molecules remain in the first, thus decreasing entropy.

[44] It could easily be argued that this same mechanism – i.e. setting precision values so that prediction errors induce no response – underpins repression (see Solms, 2018a). This is what Freud's notion of repression as "a failure of translation" amounts to within the present framework (Letter to Fliess, December 6, 1896; Freud, 1950 [1892-99]).

and it thereby determines all of its active, subjectively embodied engagement with it.

None of this can go on in the dark.

Introspective precision is inherently about selfhood and intentionality (and therefore survival). Its compulsive quality is gradually diluted as the centrifugal processing hierarchy is traversed, through instinctual and sensory affective mechanisms, and the non-declarative behavioral stereotypes associated with them, via the declarative LTM systems, to the ever-changing STM periphery (see section 'Consciousness Arises Instead of a Memory-Trace' below).

The affective value implicit in ω must be an inherent property of any self-organizing system that proactively and contextually resists the Second Law of thermodynamics. Precision optimization determines the extent to which this value will be felt (i.e., expressed via selective enabling of belief updating) for purposes of choice. To be clear: it is easy to envision an organism (or machine) in which precision values are set in such a way that the system's responses to prediction error are automatized. Indeed, large swathes of the human nervous system (not to mention the rest of the body) are organized in this way.

It is noteworthy that qualitative fluctuations in affect (i.e., ω) arise continuously from periodic comparisons between the sensory states that were predicted—based upon a generative model of the internal body ($Q\eta$) and the world ($\psi(Q)$ and samples of the actual sensory states (ϕ). This recurrent assessment of sensory states only gives rise to changes in subjective quality when the amplitude of prediction errors *changes*—signaling a change in uncertainty about the state of affairs and, in particular, the expected consequences of action (M). For this reason alone, it must be said—as one of my reviewers helpfully asked me to clarify—the Nirvana that the ideal self-organizing system described here strives for can never be attained in a real biological system, for the simple reason that change (both external and internal) always happens[45].

Below, we will briefly consider the relation of this capacity to neural plasticity. It is difficult to conceive of a complex self-organizing system adapting flexibly to changing and novel environments in the absence of some such capacity. This, in my view, is how and why consciousness arises.

"CONSCIOUSNESS ARISES INSTEAD OF A MEMORY-TRACE"

This section will be disproportionately short (see Solms, 2018a,d, in press, for fuller treatments).

We saw above that conscious self-states are fundamentally affective states. Consciousness—in its most elementary form—is a sort of alarm mechanism, which guides the behavior of self-organizing systems as they negotiate situations beyond the bounds of their preferred states, in so far as they are not equipped with automatized (or automatable) predictions for dealing with them.

I explained in section 'In the Beginning Was the Affect' that the predictions which return us complex organisms to our preferred states are provided, in the first instance, by instinctual behaviors—which are innate survival tools. These tools serve us well, and are utilized willy nilly, but they cannot possibly do justice to the complexities of the environmental niches we actually find ourselves in. For this reason, innate predictions must be supplemented through learning from experience.

That is why we *feel* instinctual emotions: we feel them because they do not and cannot predict all the variance. What we feel, in short, is the residual prediction error and associated uncertainty as we surf unpredicted situations. This (feeling within a particular modality) guides the choices which—over time—generate new, acquired predictions, in the manner described in section 'In the Beginning Was the Affect'.

But the ideal of such emotional learning is to automatize the acquired predictions (Some of them, such as fear conditioning, are automatized at the outset; others, like attachment bonding, are consolidated over longer time periods). Naturally, we need to forge new predictions which are at least as reliable as the innate ones, and to the extent that we achieve this (i.e., to the extent that prediction errors wane), to that extent acquired emotional predictions are automatized through consolidation, right down to the level of procedural memory systems (which are "hard to learn and hard to forget," see Squire, 2004). In this way, the acquired predictions come to resemble the instinctual ones, not only in their functional properties[46] but also in their subcortical anatomical localization.

The most important functional property of non-declarative memories is the very fact that they are non-declarative. This boils down to the fact that subcortical memory traces cannot be retrieved in the form of *images*, for the simple reason that they do not consist in cortical mappings of the sensory-motor surface organs[47]. They entail simpler cause-and-effect links of the kind that were described above as "associative learning of the connection between actions and their effects."

The cortical (declarative) memory systems, by contrast, are always ready, on the basis of prediction errors, to revive the mental images they represent. In other words, declarative systems readily return LTM traces to the STM state of *conscious* working memory—in order to update them[48]. This necessarily entails activation (i.e., selection) of salient cortical representations—their salience being determined (and

[45] As the Talking Heads song poetically tells us: "Heaven is a place / where nothing ever happens."

[46] Cf. (Freud's, 1915a) "special characteristics of the system Ucs," all of which can be reduced to the functional characteristics of the procedural and emotional memory systems (see Solms, 2018d).

[47] Cf. (Freud's, 1923) notion of the "bodily ego" being derived from cortical projections of the sensory-motor periphery.

[48] This property of declarative LTMs coincides exactly with what Freud called the system Pcs, although in my view the Pcs consists in both word and thing presentations (both semantic and episodic traces). Surely, there are no thing presentations in the Ucs (in non-declarative memory), only stereotyped action programmes.

"flavored") by the relevant prediction errors and variance. This process (which Friston calls "surprise") should not be confused with the sensory affect of surprise. The felt affect in question may be *any* of the homeostatic, emotional or sensory affects.

It is important to note that felt affects typically incorporate both the selected error signal *and* the ensuing adjustment of cortical (and over longer time frames, subcortical) precisions. But as the latter (cognitive) component of predictive-work-in-progress binds the former (affective) free energy, so the conscious states in question will resemble conscious thinking rather than feeling[49]. Even conscious thinking requires the presence of a subject of experience, but the process becomes unconscious just as soon as it possibly can. This coincides neatly with the fact that feeling only persists (is only required) for as long as the cognitive task at hand remains unresolved. Conscious cognitive capacity is an extremely limited resource (cf. Miller's law, above) which must be used sparingly.

In these few words, we have explained the conscious part of cognition—the part that is left over "when the performance of all the [other] functions is explained" (Chalmers).

It is hopefully clear from the foregoing that the essential task of cognitive (cortical) consciousness is to *delay* motor responses to affective "demands made upon the mind for work[50]." This delay enables thinking. The essential function of cortex is thus revealed to be stabilization of non-declarative executive processes—thereby raising them to a higher "cathectic level" (i.e., the bound state)—which is the essence of what we call (for good reason) *working* memory.

The above-described reversal of the consolidation process (*reconsolidation*; Nader et al., 2000) renders LTM-traces labile, through literal dissolution of the proteins that initially "wired" them (Hebb, 1949). This iterative feeling and re-feeling one's way through declarable problems is—on the proposal presented here—the function of the cognitive qualia which have so dominated contemporary consciousness studies. In short, conscious reconsolidation is predictive-work-in-progress. One is reminded of Freud (1920) obscure dictum: "consciousness arises instead of a memory-trace" (i.e., a labile trace is not a trace, it is a state of what Freud called drive "discharge"; see Solms, 2015).

Perceptual/cognitive consciousness (activated via attention), no less than affect itself, is a product of uncertainty. Non-declarative (subcortical) memory-traces are far less uncertain—more precise but also less complex—than declarative (cortical) ones. The relative degree of precision typically attaching to cortical vs. subcortical vs. autonomic prediction errors, therefore, coincides with the relative plasticity (resistance to change) of their associated beliefs.

One need only add that the exteroceptive sensory-motor modalities are "flavored" by consciousness in just the same way as interoceptive ones are, and for the same reason. This facilitates compartmentalization of the relevant data (and thereby reduces computational complexity) while the self-system surfs uncertainty in contextually variable conditions (The role of precision weighting in these conditions, in relation to the various perceptual modalities, and—most interestingly—in relation to language and inner speech, are discussed at length by Hohwy, 2013 and Clark, 2016).

These laconic formulations provide the basis for a new, integrated theory of affective and cognitive consciousness (and the unconscious).

CONCLUSION

In this paper, I have drawn attention to two impediments to solving the "hard problem" of consciousness—one philosophical and one scientific—and I have suggested how these impediments might be removed. The first is the popular idea that the brain "produces" consciousness, i.e., that physiological processes literally *turn into* experiences, through some curious metaphysical transformation. The second impediment is the conventional notion that consciousness is a function of cerebral cortex, i.e., that visual awareness (or any other form of conscious cognition) serves as the model example of consciousness. Adopting a dual-aspect monist position on the philosophical mind/body problem allows us to find the causal mechanism of consciousness not in the manifest brain but rather in its *functional organization*, which ultimately underpins both the physiological and the psychological manifestations of experience. In order to transcend the figurative language of dualism, this unifying (monist) organization should be described in *abstract* terms (i.e., neither in physiological nor psychological terms but rather in mathematical ones). 'Against this background,' I (like Damasio and others) suggest that the long-sought mechanism of consciousness is to be found in *an extended form of homeostasis*, which describes the mode of functioning of both the deep brainstem nuclei that provide the NCC of affective arousal and the experience of feeling itself (which appears to be the foundational form of consciousness). This type of homeostasis (formalized here as free-energy minimization) entails the generation of affects (formalized as homeostatic prediction errors) which must be contextually prioritized in relation to each other and not-system events (formalized as precision weighting), leading to modulation of perception and action (formalized as error correction) on the basis of felt uncertainty. This modulatory arousal process, in turn, leads to *learning from experience* through reconsolidation, which bestows an enormous adaptive advantage over simpler types of homeostasis—such as those found in autonomic (involuntary) nervous systems and refrigerators—the advantage being a capacity for life-preserving intentional behavior in unpredicted situations.

[49]This corresponds roughly to Freud's distinction between freely mobile and bound cathexes. However, we should not overlook the fact that the *goal* of thinking is automatization. Bound cathexes are, in short, merely *tolerated* by the ego (cf. Freud's compromise "constancy principle"). The ego's *ideal* state remains Nirvana (a curious state in which there is no residual free energy and precision becomes infinite).

[50]This coincides exactly with Freud's notion of "secondary process." Freud described the distinction between free and bound nervous energy as his "deepest insight" and added: "I do not see how we can avoid making it." (Freud, 1915a, p. 188)

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

I would like to thank Karl Friston for valuable revisions of section 'To Be Precise'.

REFERENCES

Bargh, J., and Chartrand, T. (1999). The unbearable automaticity of being. *Am. Psychol.* 54, 462–479. doi: 10.1037/0003-066X.54.7.462

Berridge, K. (1996). Food reward: brain substrates of wanting and liking. *Neurosci. Biobehav. Rev.* 20, 1–25. doi: 10.1016/0149-7634(95)00033-B

Berridge, K., Robinson, T., and Aldridge, J. W. (2009). Dissecting components of reward: 'liking', 'wanting', and learning. *Curr. Opin. Pharmacol.* 9, 65–73. doi: 10.1016/j.coph.2008.12.014

Blomstedt, P., Hariz, M., Lees, A., Silberstein, P., Limousin, P., Yelnik, J., et al. (2007). Acute severe depression induced by intraoperative stimulation of the substantia nigra: a case report. *Parkinsonism Relat. Disord.* 14, 253–256. doi: 10.1016/j.parkreldis.2007.04.005

Brentano, F. (1874). *Psychology From an Empirical Standpoint*. London: Routledge.

Carhart-Harris, R., and Friston, K. (2010). The default mode, ego functions and free energy: a neurobiological account of Freudian ideas. *Brain* 133, 1265–1283. doi: 10.1093/brain/awq010

Chalmers, D. (1995). Facing up to the problem of consciousness. *J. Consciousness Stud.* 2, 200–219.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Clark, A. (2016). *Surfing Uncertainty*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780190217013.001.0001

Craig, A. D. (2009). How do you feel – now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10, 59–70. doi: 10.1038/nrn2555

Crick, F. (1994). *The Astonishing Hypothesis*. New York, NY: Scribner.

Damasio, A. (1994). *Descartes' Error*. New York, NY: Putnam.

Damasio, A. (1999). Commentary to Panksepp, emotions as viewed by psychoanalysis and neuroscience. *Neuropsychoanalogy* 1, 38–39. doi: 10.1080/15294145.1999.10773242

Damasio, A. (2010). *Self Comes to Mind*. New York, NY: Pantheon.

Damasio, A. (2018). *The Strange Order of Things*. New York, NY: Pantheon.

Damasio, A., and Carvalho, G. (2013). The nature of feelings: evolutionary and neurobiological origins. *Nat. Rev. Neurosci.* 14, 143–152. doi: 10.1038/nrn3403

Damasio, A., Damasio, H., and Tranel, D. (2012). Persistence of feeling and sentience after bilateral damage of the insula. *Cereb. Cortex* 23, 833–846. doi: 10.1093/cercor/bhs077

Du Bois-Reymond, E. (1918). *Letter to Hallmann, 1842. Jugendbriefe von Emil Du Bois-Reymond an Eduard Hallmann*. Berlin: Dietrich Reimer. p. 108.

Ellis, G., and Solms, M. (2018). *Beyond Evolutionary Psychology: How and Why Neuropsychological Modules Arise*. Cambridge: Cambridge University Press.

Freud, S. (1893). *Charcot, Standard Edn. Vol. 3*. London: Hogarth Press. p. 11–23.

Freud, S. (1894). *The Neuro-Psychoses of Defence*, *Standard Edn. Vol. 3*. London: Hogarth Press. p. 45–61.

Freud, S. (1900). *The Interpretation of Dreams, Standard Edn*. London: Hogarth Press, 4 and 5.

Freud, S. (1901). The Psychopathology of Everyday Life. *Standard Edn*. London: Hogarth Press. p. 6.

Freud, S. (1915a). *The Unconscious*, *Standard Edn*. Vol. 14. London: Hogarth Press. p. 166–204.

Freud, S. (1915b). *Instincts and Their Vicissitudes*, *Standard Edn*. Vol. 14. London: Hogarth Press. p. 117–140.

Freud, S. (1916-17) Introductory Lectures in Psychoanalysis, *Standard Edn*. London: Hogarth Press. p. 15–16.

Freud, S. (1917). Metapsychological Supplement to the Theory of Dreams. *Standard Edn*. Vol. 14. London: Hogarth Press. p. 222–235.

Freud, S. (1920). *Beyond the Pleasure Principle*, *Standard Edn. Vol. 18*. London: Hogarth Press. p. 7–64.

Freud, S. (1923). *The Ego and the id. Standard Edn. Vol. 19*. London: Hogarth Press. p. 12–59.

Freud, S. (1925a). *An Autobiographical Study, Standard Edn*. London: Hogarth Press. p. 20.

Freud, S. (1925b). *A Note Upon "the mystic writing-pad," Standard Edn. Vol. 16*. London: Hogarth Press. p. 227–232.

Freud, S. (1940). *An Outline of Psychoanalysis*, *Standard Edn. Vol. 23*. London: Hogarth Press. p. 144–207.

Freud, S. (1950 [1892-99]). *Extracts From the Fliess Papers*, *Standard Edn. Vol. 1*. London: Hogarth Press. p. 174–280.

Freud, S. (1950 [1895]) *Project for a Scientific Psychology*. *Standard Edn. Vol. 1*. London: Hogarth Press. p. 281–397.

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475. doi: 10.1098/rsif.2013.0475

Golaszewski, S. (2016). Coma-causing brainstem lesions. *Neurology* 87:2433. doi: 10.1212/WNL.0000000000003417

Groddeck, G. (1977). *The Meaning of Illness: Selected Psychoanalytic Writings Including His Correspondence With Sigmund Freud*. London: Hogarth.

Harlow, J. M. (1868). Recovery from the passage of an iron bar through the head. *Publicat. Massachusetts Med. Soc.* 2, 327–347.

Hebb, D. (1949). *The Organization of Behavior*. New York, NY: John Wiley.

Hobson, J. A. (2009). REM sleep and dreaming: towards a theory of protoconsciousness. *Nat Rev Neurosci.* 10, 803–813. doi: 10.1038/nrn2716

Hobson, J. A., and Friston, K. (2014). Consciousness, dreams, and inference: the Cartesian theatre revisited. *J. Consciousness Stud.* 21, 6–32.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199682737.001.0001

Huston, J., and Borbely, A. (1974). The thalamic rat: general behaviour, operant learning with rewarding hypothalamic stimulation, and effects of amphetamine. *Physiol. Behav.* 12, 433–448. doi: 10.1016/0031-9384(74)90121-8

Kandel, E., Schwartz, J., Jessell, T., Siegelbaum, S., and Hudspeth, A. (2012). *Principles of Neural Science, 5th Edn*. New York, NY: Elsevier.

Kihlstrom, J. (1996). "Perception without awareness of what is perceived, learning without awareness of what is learned," in *The Science of Consciousness: Psychological, Neuropsychological and Clinical Reviews,* ed, M. Velmans (London, Routledge), p. 23–46.

Koch, C. (2004). *The Quest for Consciousness*. New York, NY: WH Freeman.

LeDoux, J., and Brown, R. (2017). A higher-order theory of emotional consciousness. *Proc. Natl.Acad. Sci. U.S.A.* 114, E2016–E2025. doi: 10.1073/pnas.1619316114

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pac. Philos. Q.* 64, 354–361. doi: 10.1111/j.1468-0114.1983.tb00207.x

Merker, B. (2007). Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behav. Brain Sci.* 30, 63–134. doi: 10.1017/S0140525X07000891

Meyer, J., and Quenzer, L. (2005). *Psychopharmacology: Drugs, The Brain, and Behavior*. Sunderland, MA: Sinauer Associates.

Meynert, T. (1884). *Psychiatrie. Klinik der Erkrankungen des Vorderhirns, begründet auf dessen Bau, Leistungen und Ernährung*, Vienna: Wilhelm Braumüller.

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. *Science* 319, 1543–1546. doi: 10.1126/science.1150769

Moruzzi, G., and Magoun, H. (1949). Brain stem reticular formation and activation of the EEG. *Electroencephalog. Clin. Neurol.* 1, 455–473. doi: 10.1016/0013-4694(49)90219-9

Munk, H. (1878). Weitere Mittheilungen zur Physiologie der Grosshirnrinde. *Arch für Physiol.* 2, 162–177

Munk, H. (1881). *Ueber die Funktionen der Grosshirnrinde*. Berlin: August Hirschwald.

Nader, K., Schafe, G. E., and Le Doux, J. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* 406, 722–726. doi: 10.1038/35021052

Nagel, T. (1974). What is it like to be a bat? *Philos. Rev.* 83, 435–450. doi: 10.2307/2183914

Panksepp, J. (1998). *Affective Neuroscience*. Oxford: Oxford University Press.

Panksepp, J., Lane, R. D., Solms, M., and Smith, R. (2016). Reconciling cognitive and affective neuroscience perspectives on the brain basis of emotional experience. *Neurosci. Biobehav. Rev.* 76, 187–215. doi: 10.1016/j.neubiorev.2016.09.010

Parvizi, J., and Damasio, A. (2003). Neuroanatomical correlates of brainstem coma. *Brain* 126, 1524–1536. doi: 10.1093/brain/awg166

Penfield, W., and Jasper, H. (1954). *Epilepsy and the Functional Anatomy of the Human Brain*. Oxford: Little & Brown.

Pfaff, D. (2006). *Brain Arousal and Information Theory*. Cambridge, MA: Harvard University Press. doi: 10.4159/9780674042100

Pulsifer, M., Brandt, J., Salorio, C., Vining, E., Carson, B., and Freeman, J. (2004). The cognitive outcome of hemispherectomy in 71 children. *Epilepsia* 45, 243–254. doi: 10.1111/j.0013-9580.2004.15303.x

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues Clin. Neurosci.* 18, 23–32.

Searle, J. (1997). *The Mystery of Consciousness*. New York, NY: New York Review of Books.

Searle, J. (2017). "Foreword," in *Biophysics of Consciousness: A Foundational Approach. eds* R. Poznanski, J. Tuszynski and T. Feinberg (New York, NY: World Scientific) 129–148.

Sharma, J., Angelucci, A., and Sur, M. (2000). Induction of visual orientation modules in auditory cortex. *Nature* 404, 841–847. doi: 10.1038/35009043

Shewmon, D., Holmse, D., and Byrne, P. (1999). Consciousness in congenitally decorticate children: developmental vegetative state as a self-fulfilling prophecy. *Dev. Med. Child Neurol.* 41, 364–374. doi: 10.1017/S0012162299000821

Solms, M. (1997). What is consciousness? [and response to commentaries]. *J. Amer. Psychoanal. Assn.* 45, 681–778. doi: 10.1177/000306519704500 31201

Solms, M. (2013). The conscious id. [and response to commentaries]. *Neuropsychoanalysis* 15, 5–85 doi: 10.1080/15294145.2013.10773711

Solms, M. (2014). A neuropsychoanalytical approach to the hard problem of consciousness. *J. Integr. Neurosci.* 13, 173–185. doi: 10.1142/S0219635214400032

Solms, M. (2015). Reconsolidation: turning consciousness into memory. *Behav. Brain Sci.* 38, 40–41. doi: 10.1017/S0140525X14000296

Solms, M. (2018a). What is 'the unconscious' and where is it located in the brain? *Ann. NY Acad. Sci.* 1406, 90–97.

Solms, M. (2018b). Review of damasio, the strange order of things. *J. Am. Psychoanal. Ass.* 66, 579–586. doi: 10.1177/0003065118780182

Solms, M. (2018c). Extracts from the revised standard edition of Freud's complete psychological works. *Int. J. Psychoanal.* 99, 11–57. doi: 10.1080/00207578.2017.1408306

Solms, M. (2018d). The neurobiological underpinnings of psychoanalytic theory and therapy. *Front. Behav. Neurosci*. 12:294. doi: 10.3389/fnbeh.2018.00294

Solms, M. (in press). *Consciousness Itself: Feeling and Uncertainty*. London: Profile Books.

Solms, M., and Friston, K. (2018). How and why consciousness arises: some considerations from physics and physiology. *J. Conscious. Stud.* 25, 202–238.

Solms, M., Kaplan-Solms, K., and Brown, J. W. (1996). "Wilbrand's case of 'mind blindness,'" in *Classic Cases in Neuropsychology*, eds C. Code, Y. Joanette, A. Lecours, and C.-W. Wallesch (London: Psychology Press), 89–110.

Solms, M., and Panksepp, J. (2012). The 'id' knows more than the 'ego' admits. *Brain Sci.* 2, 147–175. doi: 10.3390/brainsci2020147

Solms, M., and Saling, M. (1990). *A Moment of Transition: Two Neuroscientific Articles by Sigmund Freud*. London: Karnac.

Squire, L. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiol. Learn. Mem.* 82, 171–177. doi: 10.1016/j.nlm.2004.06.005

Strachey, J. (1962). *The Emergence of Freud's Fundamental Hypotheses. Standard Edn* (London: Early Psycho-Analytic Publications), 3, 62–68.

Teuber, H.-L. (1955). Physiological psychology. *Ann. Rev. Psychol.* 6, 267–296. doi: 10.1146/annurev.ps.06.020155.001411

Tononi, G. (2012). *Phi: A Voyage from the Brain to the Soul*. New York, NY: Pantheon.

Wakefield, J. (2018). *Freud and Philosophy of Mind, Volume 1: Reconstructing the Argument for Unconscious Mental States*. London: Palgrave Macmillan doi: 10.1007/978-3-319-96343-3

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 Psychology

HYPOTHESIS AND THEORY
published: 28 March 2019
doi: 10.3389/fpsyg.2019.00592

"Surprise" and the Bayesian Brain: Implications for Psychotherapy Theory and Practice

Jeremy Holmes[1] and Tobias Nolte[2]*

[1]University College London, Anna Freud National Centre for Children and Families, London, United Kingdom, [2]Department of Psychology, College of Life and Environmental Sciences, University of Exeter, Exeter, United Kingdom

The free energy principle (FEP) has gained widespread interest and growing acceptance as a new paradigm of brain function, but has had little impact on the theory and practice of psychotherapy. The aim of this paper is to redress this. Brains rely on Bayesian inference during which "bottom-up" sensations are matched with "top-down" predictions. Discrepancies result in "prediction error." The brain abhors informational "surprise," which is minimized by (1) action enhancing the statistical likelihood of sensory samples, (2) revising inferences in the light of experience, updating "priors" to reality-aligned "posteriors," and (3) optimizing the complexity of our generative models of a capricious world. In all three, free energy is converted to bound energy. In psychopathology energy either remains unbound, as in trauma and inhibition of agency, or manifests restricted, anachronistic "top-down" narratives. Psychotherapy fosters client agency, linguistic and practical. Temporary uncoupling bottom-up from top-down automatism and fostering scrutinized simulations sets a number of salutary processes in train. *Mentalising* enriches Bayesian inference, enabling experience and feeling states to be "metabolized" and assimilated. "Free association" enhances more inclusive sensory sampling, while dream analysis foregrounds salient emotional themes as "attractors." FEP parallels with psychoanalytic theory are outlined, including Freud's unpublished project, Bion's "contact barrier" concept, the Fonagy/Target model of sexuality, Laplanche's therapist as "enigmatic signifier," and the role of projective identification. The therapy stimulates patients to become aware of and revise the priors' they bring to interpersonal experience. In the therapeutic "duet for one," the energy binding skills and non-partisan stance of the analyst help sufferers face trauma without being overwhelmed by psychic entropy. Overall, the FEP provides a sound theoretical basis for psychotherapy practice, training, and research.

Keywords: Bayesian brain, psychoanalysis, active inference, psychotherapy, free energy principle, mentalization

OPEN ACCESS

Edited by:
Jim Hopkins,
University College London,
United Kingdom

Reviewed by:
Patrick Connolly,
Hong Kong Shue Yan University,
Hong Kong
Tamara Fischmann,
International Psychoanalytic
University Berlin, Germany

****Correspondence:***
Jeremy Holmes
j.a.holmes@btinternet.com

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *19 November 2018*
Accepted: *04 March 2019*
Published: *28 March 2019*

Citation:
Holmes J and Nolte T (2019) "Surprise" and the Bayesian Brain: Implications for Psychotherapy Theory and Practice. Front. Psychol. 10:592. doi: 10.3389/fpsyg.2019.00592

INTRODUCTION

It has been established beyond doubt that psychodynamic psychotherapy "works" (Leichsenring, 2008; Shedler, 2010; Leichsenring et al., 2015; Taylor, 2015). But how? Building on recent advances in computational neuroscience, the aim of this paper is to offer a heuristic that can help elucidate the underlying mechanisms by which psychotherapies alleviate psychological distress and illness.[1]

[1]We believe that this attempt to elucidate the "neuronal" basis of effective psychotherapy exemplifies the normal course of scientific progress. Darwin knew no more about DNA than did Freud about the fMRI-unveiled brain.

Schröedinger (1944) coined the term "negentropy" to characterize the complexity of living matter, i.e., its structured heterogeneity and order, in contrast to the entropy of the inanimate world under the sway of the second law of thermodynamics. Our approach is based on the "free energy" (FE) principle developed by Friston as one formulation of "the Bayesian Brain" (Friston, 2010; Hobson and Friston, 2012; Hohwy, 2013; Friston and Frith, 2015; Hopkins, 2016). Friston presupposes that the brain's aim, like that of the organism as a whole, is to maintain homeostasis[2] and resist the entropic forces of chaos and homogenization. To do this we–along with our fellow living creatures–need *information* about the environment, our place within it, and the likely outcomes of our actions. The past shapes our futures: based on prior experience, we make "top down" predictions about our sensory and interoceptive input, based on a model of how they were created.[3] The discrepancy between these top-down predictions and the actuality–and accuracy–of bottom-up sensations is "prediction error." *Via* perception and action, these unavoidable "errors" are "minimized" by converting *prior* beliefs into *posteriors*[4] *(i.e., the newly assigned probability after the relevant evidence, the observed data, is taken into account)*. This process of Bayesian inference simulates past experience and ensures posterior beliefs align with newly sampled data.

Prediction error is inescapable for two reasons: first, we live in a constantly changing environment, and second, our sampling of that environment is subject to inaccuracy and misperception. But this "error" is all to the good –it is the very stuff that drives a continuous process of belief-updating and helps build adaptive models of the worlds (and bodies) we inhabit.

The free energy principle (FEP) regards creatures (like us) as self-organizing systems that resist a tendency to dissipation and disorder. This applies as much to the brain in its search for meaning (i.e., informational order) as it does to the body as a whole in its pursuit of physical structure and regulation (Friston, 2013). This informational slant on "entropy" equates to "surprise."[5] If an event is probable to a high degree, the surprise when it occurs is minimal and thus little new information is gained. We can therefore be regarded as creatures that place an upper bound on free energy by minimizing their surprise, or maximizing the evidence for their models of the world. This is sometimes known as self-evidencing (Hohwy, 2016). Free energy can be decomposed into *complexity* minus *accuracy*. Accuracy refers to our ability to predict sensations, while complexity reflects the degrees of freedom used to provide an accurate prediction. Model evidence increases by minimizing free energy. The accuracy of predictions rises, with a "concomitant increase in complexity so that increased model complexity is always licensed by an ability to make more accurate predictions" (Solms, 2019).

This *predictive coding visualises* the brain as engaged in neuronal–and, as we shall argue, conceptual–dynamics, that minimize free energy by working to reduce prediction errors. The latter are the difference between sensory input and predictions of that input based upon expectations about states of the world created by a pre-existing "generative model." Resolving prediction errors updates prior beliefs by converting them into posterior beliefs. The empirical evidence from neuroscience suggests that this process rests upon (forward or "bottom-up") prediction errors that ascend brain hierarchies from the low sensory levels to high levels of deep generative models (Carhart-Harris and Friston, 2010). For example, the number of "top-down" efferent neurons targeting the eye far exceeds the "bottom-up" afferent number ascending brain-ward. Descending predictions try to resolve prediction errors at each hierarchical level, thereby providing an accurate account of sensations, in a minimally complex fashion.

The FEP provides a model to think about belief updating and what this might entail. The binding of free energy equates to the resolution of prediction errors (i.e., surprise and uncertainty). Thus, the conversion of free into bound energy results from belief-updating to increase the accuracy–or decrease the complexity–associated with our beliefs about the world's states of affairs.

In sum, Friston maintains that the brain's main aim is to minimize "surprise"–as best it can.

Prediction error is minimized in two main ways:

1. *Action*, which reduces prediction errors by selectively sampling sensations that are the least surprising,[6] thereby helping to approximate the organism to its environmental niche, or affordance (see below).
2. *Perception*. Changed perceptions follow from belief updating resulting in more reality-consonant *predictions*.

Both action and perception operate semi-instantaneously–in the twinkling of an eye. In the longer term, the structure of generative models are, in health, continuously being updated, especially their *complexit*y. How this plays out in psychopathology are main themes of this article. Much of our focus will be on what Friston and collaborators call "structure learning" (Tervo et al., 2016; Friston et al., 2017; Gershman, 2017; Isomura and Friston, 2018), namely, learning the repertoire or narratives that constitute our prior beliefs–or hypotheses–about how our world works, and how these might be influenced therapeutically. Although the FEP applies to these structural priors, getting them right can be a tricky business. If we have too many

[2]Sterling (2012) has introduced the term "allostasis" to capture a more dynamic version of homeostasis in which an organism anticipates change in the internal milieu and sets about counteractive processes and actions.
[3]A concept that can be aligned with the psychoanalytic notion of "repetition compulsivity" (Barratt, 2016), or, more poetically, Wordsworth's child as father of the man.
[4]The terms derive from Kant's *a priori* and *a posteriori*.
[5]Surprise is defined as the *negative log-probability of an outcome*, i.e., how "likely" or "unlikely" a particular event, from a specific organism's viewpoint to occur. The brain cannot compute "surprise" as such, but free energy *can* be evaluated and by "active inference." Active inference depends on two key processes: modifying sensory input "bottom-up" from sensory epithelia, including the interoceptive, affect-triggering receptors, (Barrett, 2017), and "top-down" from the cortex – and at intermediate levels in between.
[6]This a key point of intersection between Bayesian predictive processing theories and "embodied enactive" models of the mind which prevail in cognitive science (Hohwy, 2013; Kirchhoff, 2017).

prior hypotheses, our models are too complex and will not generalize in a capricious and changing world. Conversely, if we have overly simplistic models, with an insufficient number of priors to call upon we will fail to predict our sensations accurately. In both cases, free energy increases and we fail as self-evidencing creatures. We shall argue that psychopathology largely resides in the discrepancy between the experience of uncertainty and the paucity or defectiveness of procedures needed to reduce it.

It is important to note that minimizing surprise does not equate to stasis or clinging to the *status quo*. First, the internal milieu, i.e., physiology is constantly changing and so interoceptive prediction error will drive appetitive, safety-seeking, and reproductive behaviors (Seth, 2015) exploiting an innate system whose prediction postulates that "engagement with a source of uncertainty provides maximal opportunities to resolve that uncertainty" (Solms, 2019). Second, organisms live in constantly changing environments, both in the short- and long term, and need creative solutions to adjust and adapt to these. Integral to this is the invisible and imperceptible flux of time. This aspect of active inference can be thought of in terms of "time out" *simulations*. By uncoupling prediction and action, the mind models the possible outcomes of action in terms of expected surprise or uncertainty. Thus active inference furnishes building blocks for allostatic adjustment, i.e., "flexible information manipulation without the need to commit to particular decisions at an early stage of processing" (Knill and Pouget, 2004). Seen this way, imaginative exploration and innovation are no less surprise-minimizing than ingrained, self-perpetuating, ways of explaining the lived world. It is this former aspect that is built on and prosthetically enhanced in the social practices of psychotherapy.

PSYCHOANALYTIC RESONANCES

At first encounter, this abbreviated account may seem to come from a conceptual universe far removed from psychoanalysis. Knowing our left from our right hand,[7] active inference can no doubt reliably discount the chances of a west-rising dawn. But knowing about the physics of the world "out there" is surely a very different matter to the task of understanding oneself and other people? The argument of this paper is, to the contrary, that Fristonian principles apply equally, if not more so, to the interpersonal realm.

Consider a baby crying for its mother. At times, she is there on demand; at others, she is inexplicably delayed. In order to make good predictions, a theory of mind is needed–"maybe she's tired, angry about my neediness, intoxicated, making a new potential rival with Dad."

The Bayesian brain gradually–and with help–learns to infer the causes, affects, motivations, and meanings which shape the interpersonal world. Prediction error is built into this calculus; this will steer *actions*, aiming to minimize expected error and therefore, *via belief updating*, increase the chances of our predictions being adaptively correct:

> "Mummy, I called you last night when I had a tummy ache, but you didn't come! I thought you had gone away"
> "So sorry darling, how horrid! I must have been fast asleep. If it happens again you must come through and wake me up." (c.f., Allen et al., 2018)[8]

Here, the child is being taught the role of action ("come through"), interoceptive affect regulation ("So sorry–how horrid"), and a relevant hypothesis or prior ("maybe she's asleep and can't hear me"). Note the *conversational* or narrative aspect of prior/posterior interplay. *Vis-a-vis* the physical world, action is used to minimize the discrepancy between the organism's model and environmental "affordances" (Dennett, 2017) that themselves can be purely epistemic–in the sense of resolving surprise and uncertainty. In the interpersonal world, dialogue is not so much with physical objects–moving one's head to get a better view, etc.–but with the other, engaged in a reciprocal project of speech acts (c.f., Talia et al., 2014). At this level of the Bayesian hierarchy, prior beliefs are higher order cognitions (HOCs; Rudrauf, 2014; Debbané and Nolte, 2019), initially "borrowed" by infants from parents' minds, based upon their caregiving disposition. We shall see how similar processes apply to psychoanalytic work.

This moves the Free Energy approach toward developmental and interpersonal conceptions with which psychoanalysis can begin to engage. Consider three relevant aspects. First, when it comes to precedence in the concept of free energy, Freud trumps Friston (Cahart-Harris and Friston, 2010; Solms, 2013). In the unpublished "Project" Freud (1895/1950) proposed the concepts of "Bindung" and "Entbindung," i.e., energy "bound" and unbound."[9] Freud abandoned his "project," as he moved toward more psychological models of the mind. However, in his 1911 paper *Formulations on the Two principles of Mental Functioning*, he differentiates primary process thinking, in which libido seeks discharge, from secondary processes which encompass language, sublimation, and ego-mediated restraint. The primary processes can be thought of as bottom-up impulses (interoceptions) stimulating and interacting with the top-down secondary process of affective modulation, verbal representation, and logic. For Freud the aim is homeostasis or psychic equilibrium, through binding, or if that fails, "discharge" in form of symptoms:

> "The purpose of the mental apparatus [is] to keep as low as possible the total amount of the excitations to which it is subject" (Freud, 1925).

Relevant to our later discussion of trauma is emphasis on painful memories, which, if unregulated, remain disruptively

[7]Many metrics, affective and cognitive, start from the body orientation (Lakoff and Johnson, 2003).

[8]For a recent example of a simulated infant learning about mother's quality of caregiving under active inference, see Cittern et al. (2018).

[9]Freud, well versed in classical literature, would have been familiar with Aeschylus' play *Prometheus Bound*, and perhaps with Shelley's subversive version of the myth, *Prometheus Unbound*.

unbound (Freud, 1895/1950). On the free energy view, this corresponds to unresolved surprise, uncertainty, or prediction errors–all which may be experienced as mental pain, and therefore part of the terrain of psychoanalytic therapy.

Another close parallel is between Friston's model and Bion's (1962) quasi-mathematical picture of how alpha function (i.e., maternal reverie generating top-down predictions) processes infants' "beta elements" (uncontained, unnamed bottom-up raw experience) (c.f., Mellor, 2018). This "borrowed brain" (Holmes and Slade, 2017) model introduces a vital interpersonal dimension to the Bayesian process. Parental mentalizing–seeing, understanding, and resonating with their infants' affects–is initially non-verbal and implicit: communicated by facial expression, tone of voice, affiliative touch, swinging rhythms of soothing, or stimulation. These embodied gestures present a model of the infant from the caregiver's perspective, helping the child to integrate primary sensory signals (Fotopoulou and Tsakiris, 2017) into regularities of emotional and interpersonal consequences. In the context of increasing predictability, the infant explores the environment (beginning with the mother's breast) and the mind of others with unconscious phantasies and proto-representations (i.e., building a repertoire of Bayesian "priors"). With the help of predictable input from the caregiver, the infant brain begins to differentiate self versus non-self causes of sensations that underwrite a sense of agency and the emergence of selfhood (Fonagy et al., 2002).

This leads us to a third Friston-Freud link: the analysis of *boundaries*. Bion postulated a "contact barrier" between conscious and unconscious thought, ensuring that phantasy is sharply differentiated from reality–the pleasure from the reality principles, the gratifying from the missing–and much-missed–breast.[10]

Comparably, from a FEP perspective, living entities possess a statistically permeable boundary across which occur exchanges–material and informational–with their surroundings. The mind is *bounded*; at one level, the "world" can only be known *via* its impression on the sensory epithelium and the belief updating entailed by active inference that sensations evoke. This boundary (known as a Markov blanket, see Kirchhoff, 2017; Kirchhoff et al., 2018) demarcates any system or creature from the environment in which it is immersed and also describes nested layers of top-down/bottom-up interfaces within the brain.

The "world" is opaque to the brain except insofar as it samples sensations from outside across the Markov blanket, matching them with its own internally generated models, identifying discrepancies as prediction errors and acting and/or thinking to minimize them. As seen (felt, smelled, heard, propriocepted) through a Markov blanket, "the world" is inferred, based on sensation: seeing, feeling, etc. is believing. Markov blankets are "nested," in the sense that boundaries exist not just between the mind and its environment, but within the body-mind at different levels of complexity and immediacy. Bottom-up and top-down processes interact in a hierarchical Helmholtzean fashion throughout the nervous system. Thus, believing is also seeing.

Another connection between FEP and the preoccupations of psychotherapy is the role of the self. From a FEP perspective, the "inner world"–bounded and entropy-defying –necessarily entails a model of the environment (Conant and Ashby, 1970)[11] and the organism's place within it. This presupposes a rudimentary "self" however primitive or unconscious that representation might be.[12] Enhancing the sense of self–active, authentic, aware, and apposite–is a key aim of psychotherapy.

BAYES IN ACTION

Let's now look now at a quotidian example illustrating the Bayesian brain in action, and its relevance to psychotherapy.

> One spring morning, in the course of JH's daily run across agricultural land, he noticed that the farmer had recently sprayed weed-killer. As he ran, he experienced an unpleasant sickly smell and slight feeling of nausea. Worried that he might be adversely affected, as he had been in previous years, he returned *via* a detour. The following day, following the same course, the smell had gone, but he noted in his *peripheral vision* a dark flapping object. His first thought was that this was a bird, perhaps a crow, affected by the previous day's poison; he turned his head to engage foveal/*central vision*, then *approached* to investigate further and if necessary rescue the creature. The closer he got to the "object" however, the more the putative stricken bird revealed itself to be no more than a fragment of wind-blown black plastic, a remnant of a discarded fertilizer bag.

This trivial incident illustrates a number of the Bayesian FEPs.

- JH's slight feeling of nausea on the previous day, and knowledge of the hazards of weed-spraying, raised the "prior" probability of a "sick bird." This "somatising" mind-set was based on the previous day's nausea.
- The "*prior*," or meaning attributed to this experience, based on *selective sampling* in *peripheral vision* and therefore error-prone, was guided by interoception (the feeling of sickness) and the epistemic affordance[13] of looking more closely at the cause of sensations.

[10]The latter two distinctions representing rudimentary generative models which, as unconscious phantasies, gradually become imbued with psychic meaning.

[11]See Seth (2015) for a discussion of the psychiatrist Ross Ashby's early contributions to FEP.

[12]C.f., O'Keefe (1978) who discovered "place cells" in the hippocampus which, like an internal GPS, tells mammals where they are in their world. Knowing "who" we are entails, among other information, knowing "where" we are.

[13]Gibson defines affordances as "The *affordances* of the environment are what it *offers* the [individual], what it *provides* or *furnishes*, either for good or ill… [The word affordance] implies the complementarity of the [individual] and the environment." (Gibson, 1986, p. 127). An "epistemic affordance" refers to the *meaning* of an object or event in the environment, in this case a "dark flapping object."

- The stimulus was ambiguous and, thanks to the inherent imprecision of peripheral vision, "noisy"; thus *free energy minimization* was required, *via*

    1. *Action*–turning the head and *moving toward* the "flapping" in order to disambiguate (c.f., Seth, 2015) and increase perceptual accuracy–reducing uncertainty and subsequent surprise.
    2. Belief updating–or hypothesis-revision ("the poison will have dispelled by today so it would be odd/ anomalous if this really was stricken bird").

- This *active inference*, led to a
- *Posterior belief*: a free energy-minimized explanation of reality, external ("it's only flapping plastic") and internal ("no more nausea; I'm not going to get ill").

We shall return to this example in our discussion of transference.

MENTALISING

As already mentioned, integral to active inference is an organism's "*sense of self.*" In humans and other primates, this implies the emergent property of self-awareness (Seth, 2015; Seth and Friston, 2016; Friston, 2018). The better we know who "we" are, the less likely we are to be entrapped in prediction error. Being able to model the consequences of our actions means, we have models of a counterfactual future, and thus to choose how we perceive the world and how to act on its affordances. The healthy brain is both prediction and action generator, constantly attempting to align perceived reality with internalized models (Bolis and Schilbach, 2017), including factoring in the self as a source of potential error and uncertainty. To the extent that psychotherapy helps its subjects to know themselves better, the more these processes will be enhanced.

FEP holds that nested Markov blankets operate "all the way up" (Kirchhoff et al., 2018). Thus, the search for self-awareness points to a further level of the top-down/bottom-up hierarchy (Wilson, 2002): *meta-cognition*, the capacity to think about thinking, or *mentalise* (Frith, 2012). Mentalising is the capacity to stand outside oneself and scrutinize one's–and others'–active inference. The processes by which we populate our *umwelt* with objects, motivations and meanings operate below consciousness most of the time–until problems arise, as they inevitably do; given the complexity of the social and physical environments in which humans find themselves. This is especially true of the inherently unreliable nature of self-appraisal, and the related need to navigate the shared affective world of intimate others (see Rudrauf and Debbané, 2018 for the Projective Consciousness Model of such inference processes).

Frith (2012) argues that such metacognition is especially relevant to the cooperative or "we-mode" procedures, which occupy a great deal of human waking life. He cites a range of experimental evidence showing how inaccurate unmodulated self-appraisal can be–we cannot easily see ourselves as others see us. He has shown experimentally how two heads are better than one: "through discussions of our perceptual experiences with others, we can detect sensory signals more accurately." (Frith, 2012)

Active inference, if carried out jointly, surpasses lone attempts to reduce prediction error and forestall entropic surprise. Developmental studies show how an "intimate other"–typically an attachment figure – knows our self better than we can we know ourselves, and it is through this joint appraisal that our internal self-model becomes progressively refined in the course of psychological development (Moutoussis et al., 2014; Palmer et al., 2015; Hamilton and Lind, 2016; Fotopoulou and Tsakiris, 2017). One of the roles of psychotherapy is to reactivate this process.

"Duets for One"[14]

This dyadic self slant takes us to the question: what happens when two Bayesian brains interact? Friston and Frith (2015) stake out the maths of this, using birdsong as a paradigm for dialogic "conversations." The authors base their discussion on the phenomenon of "sensory attenuation" (Brown et al., 2013), in which sensory feed-forward is inhibited during action, in order to preclude the log-jam that arises if bottom-up were to meet top-down *in medias res*.

This sensory attenuation is integral to "turn taking," as a fundamental feature of human interactions, whether verbal or non-verbal (Holler et al., 2015). One can either listen or talk, but not both. In intimate conversations one can, through the other's ears, "hear," and so come to know oneself better. If each agent assumes the other is "like" themselves, the boundaries between them are temporarily dissolved. Listening, the sensory input of A (i.e., "language," verbal and non-verbal) can be taken and "priored" (i.e., subjected to top-down predictions) as though it arose in B herself. This in turn leads to "action" (i.e., more speech), revised posteriors, and so on–a similar process applying to B *vis-a-vis* A. As Friston and Frith (2015, p. 14) put it, the result is

> "a collective narrative that is shared among communicating agents (including oneself). For example, when in conversation or singing a duet, our beliefs about the (proprioceptive and auditory) sensations we experience are based upon expectations about the song. These beliefs transcend agency in the sense that the song (e.g., hymn) does not belong to you or me"

The resulting boundary dissolving synchrony of Friston and Frith's birdsong model (i.e., "epistemic match"[15]) points the way to the nature of therapeutic conversations in psychotherapy.

[14]A phrase borrowed from Kempkinsy's play of the same name and later film, a thinly described depiction of the life and illness of the cellist Jacqueline du Pre – including the questionable role of her psychiatrist!

[15]Fonagy and Allison (2014) argue that relaxing epistemic vigilance is achieved in normal development through "prefacing" one's communicative intents with ostensive cues. This validates the recipient as a subjective, agentive self. Once epistemic trust is stimulated in this way, the channel for the transmission of knowledge – learning about minds – is opened and an *epistemic match* (Fonagy, personal communication 2018) can be created whereby one's imagined self-narrative or feeling state can be recognized in the way the other communicates their version thereof.

The therapeutic "duet for one" helps bind potentially disruptive free energy in creative ways, fostering psychological resilience. It also provides a neuroscience account of the psychoanalytic notion of the "third" (Ogden, 1994), the phantasy-imbued conversation which arises between two intimate participants (i.e., analyst and analysand), contributed to by both, but pertaining to neither.

FREE ENERGY, ATTACHMENT, AND PSYCHOPATHOLOGY

Free energy minimization describes how organisms adapt to unpredictable environments, forming a bulwark against entropy, and a springboard for survival and flourishing. But the negentropy which characterizes living organisms is inherently fragile. Given an entropic world, as the Red Queen famously puts it, "*it takes all the running you can do to stay in the same place. If you want to get somewhere you must run twice as fast….*" (Carroll, 1871/2009).[16]

This fragility, arguably, is the basis of psychological illness/psychopathology. Things can–and do–go wrong in a number of different ways (Solms, 2015; Powers et al., 2018). First, there is the ever-present danger of "trauma." Despite best laid plans, unpredictable, unforeseen, and deleterious environmental impingements can overwhelm prediction error minimization. As Freud put it:

> "we describe as 'traumatic' any excitations from outside…powerful enough to break through the protective shield…and result in permanent disturbances of the manner in which the energy operates." (Freud, 1925, p. 3)

The Markov boundaries ("blankets") of body and mind form the basis for adaptive living. The environment is "taken in" in order that it may be appraised and evaluated but also kept at bay so that it can be manipulated to the organism's advantage. The same goes for internally generated impingements, phantasies, demands, urges, or drives. In trauma, entropy, i.e., free energy unbound, takes over at a specific level of nested Markov blanket (for instance, the expectation of a safe or relatively predictable world). The mind is colonized by chaos and the potential for psychotic functioning increases if the thinking apparatus itself is overwhelmed, or as it might be put psychoanalytically, "attacked." Trauma, from this perspective, exerts pressure for parameter adjustments in generative models to deal with increased complexity that arises from traumatic experiences (Hopkins, 2016).

Second, the capacity for active inference may be impaired. Active inference, as the term implies, depends on agency and belief-updating. Both are skills, acquired and honed in the course of development and reflecting the role of caregivers, and thus vulnerable to environmental disruption. It is this acquisition that underlies structure learning and building–in a familial and an encultured setting–the right sort of priors for explaining dyadic interactions with others and our own bodies.

Seen this way, psychopathology results either from the impact of overwhelming trauma, or when the capacity for active Bayesian inference is compromised. Here, the attachment ontogenetic schema for categorizing intimate relationships provides an evidential heuristic. Insecure attachments compromise active inference (Holmes and Slade, 2017): in the absence of an internal secure base (Holmes, 2010), exploration, physical and psychological, is curtailed. This limits the extent and range of sensory sampling of the environment, and so the variety of priors or hypotheses available to account for them. Both the "breaking" (i.e., creative destruction) of priors and the "making" (i.e., creative construction) of new ones are inhibited (c.f., Holmes, 2010; Leonidaki et al., 2018).

In anxious or "hyperactivating" attachment, agency tends to be absent or eroded. Rather than actively searching or changing their environment, sufferers remain passive in the face of loss, conflict, or trauma (Knox, 2010), a state famously described as "learned helplessness" (Maier and Seligman, 2016). Here, the self is suffused with unmodulated affect. In terms of structure learning, commitment to the single prior "nothing I do will change anything" precludes epistemic affordance and the testing of alternative hypotheses. By contrast, the hallmark of deactivating, or dismissive attachments is repression and affect suppression. While this yields a measure of niche-specific security, it also renders the individual vulnerable to unexpected trauma or interpersonal friction, as well as precipitating health-diminishing physiological changes.[17]

One of the "functions" of negative affect–fear, sadness, mental pain–is as *signals* of prediction error (c.f., Barrett, 2017; Solms, 2019), i.e., a discrepancy between top-down expectation and bottom-up signal–the wanted breast and the reality of its non-appearance. If, as in anxious attachment, negative effects are felt to be un-minimizable this may lead to–or indeed constitute–psychological illness. In deactivating attachments there is a trade-off between free energy minimizing and complexity reduction. By placing interceptions beyond conscious awareness–and so beyond mentalising–the learning of adaptive structural "priors" is precluded.

Disorganized attachment is a proven precursor to later psychopathology including Borderline Personality Disorder (Bateman and Fonagy, 2012). Two main reasons have been identified. First is the low threshold for interpersonal distress typical of such individuals, which means that mentalising and so top-down modulation–free energy minimizing–of negative affect are inhibited (Nolte et al., 2013). Second, sufferers typically experience from "epistemic mistrust" (Fonagy and Allison, 2014), resulting in difficulties with the collaborative mentalising/social learning "duets" described above (Nolte et al., 2019). In such a solipsistic world, deliberate self-harm, substance abuse or risky

[16]The "Red Queen hypothesis" in evolutionary biology (Ridley, 1993) is used to account for the apparently wasteful phenomenon of sexual ("twice") as opposed to asexual reproduction.

[17]Avoidant infants separated from their care-giver appear unperturbed, but demonstrate raised cortisol and pulse rates suggestive of physiological stress (Bernard et al., 2013) with potentially long-term adverse health implications.

sex are self-soothing last resorts; however self-defeating. Bion's (1962) "minus K"–i.e., the "active," dynamically motivated wish *not* to know is also relevant. Selective sensory sampling (including interceptive input) which excludes new information means that simplistic, albeit dysfunctional models, of the world are maintained.

In all three patterns of insecure attachment, freedom is sacrificed for the sake of a degree of security. Freud defined neurosis as a turning away from reality. From a FE perspective, this can be seen in terms of attempts to bind free energy by reducing complexity. Fixed beliefs about the world are clung to, rather than updated in the light of experience. The more precision–which may be spurious–is afforded prior beliefs,[18] the less likely are new experiences sought in order to update generative models. A degree of negative capability,[19] or creative not-knowing–and hence the need for exploration and innovation–is thus built into the free energy formulation. In the Kleinian dichotomy, PSP (paranoid-schizoid position; Klein, 1946) represents a simplistic either/or good/bad model, while DP (depressive position; Klein, 1997) a more complex, whole and nuanced approximation to the world's (epistemic and affective) affordances.

Parsimony[20] plays an important role here, i.e., the need to reduce, Goldilocks fashion (neither too many nor too few), the chaotic multiplicity of possible predictions to a number of stable "attractors."[21] Such parsimonious models of the world must have value[22], i.e., be of interest to the organism, and help with its project of survival, maintaining homeostasis, facilitating consciousness, staying safe, enhancing foraging potential, reproduction, etc. Their function ultimately is to minimize the affective manifestations of chronic prediction error.

On this reading, the free energy formulation is inherently motivational. This has psychotherapeutic relevance given that therapy is ultimately concerned with people's needs, wishes, and wants (c.f., Hopkins, 2016). Moving toward complexity-reducing, parsimonious attractors that enhance interpersonal satisfaction–and eluding self-fulfilling priors (e.g., learned helplessness) are markers for psychological health.[23] Our contention is that the procedures of psychotherapy, and especially psychoanalytic variants, are well placed to enhance these processes.

HOW PSYCHOTHERAPY FOSTERS ACTIVE INFERENCE

Bio-Behavioral Synchrony Reduces "Surprise"

Bio-behavioral synchrony (Feldman, 2015a) refers to the physiological, endocrinological, and behavioral entraining characteristic of care-givers and their infants, and their developmental sequelae–and can be seen as a prototype for life-long "duets for one." The greater the synchrony in the first year of life, the more pro-social, exploratory, and less anxious the child is likely to be at school entry (Feldman, 2015b).

Bio-behavioral synchrony takes place during "sensitive periods" (Tottenham, 2014) in which immature individuals are open to affect co-regulation, with the help of their care-givers. Thus, in the classic "visual cliff" paradigm (Gibson and Walk, 1960), 1-year-old children are more adventurous and take greater risks if their mothers are seen to be encouraging and reassuring. This relational regulation is not confined to human mammals (Hofer, 2002). In the presence of their mothers, rat pups show interest in–rather than aversion to–strong odors, compared to those separated from their mothers at birth, and when mature show diminished startle reflexes and greater exploratory drive.

Secure attachment transmits epistemic trust as a springboard for social and physical exploration cross the life cycle. Coan et al. (2006) and Coan (2016) studied happily married couples in their "hand-holding" experiments. The wives were exposed to stress–the threat of a mild electric shock–while being observed in an fMRI scanner. Markers of HPA axis arousal were minimal or non-existent when holding their husbands' hands as compared with facing the threat on their own. From a free energy perspective, prediction error is lessened in these dyadic scenarios. Instead of a fast track (Kahneman, 2011), low-precision "danger" attractor, in the "duet for one" scenario, the potential free energy of threat is minimized. The "victim's" threat-induced arousal does not directly impact the hand-holding husband's HPA axis, who is thereby able to bring "top-down" reassurance into the shared experience. Undertaken together, the whole mini-trauma becomes negligible. The husband's bound energy pathways transmit the thought to his wife: "the experimenter is not really going to do anything nasty to us."[24]

Clients entering psychotherapy have typically had reduced sensitive periods of affiliative learning in their developmental histories, or, worse, attachment bonds reinforced not by collaboration and pleasure but by aversive stimuli. (Hofer, 2002). Many, especially those with a history of disorganized attachment, are on "hair trigger" for overwhelming anxiety (Allen et al., 2008). They are in the grip of perceptual distortion and ingrained

[18]Thus, OCD can be thought of in FE terms as fruitless striving for spurious certainty. In a riposte to Socrates' much-quoted aphorism that "the unexamined life is not worth living," Dennett reminds us that "the over-examined life is nothing much to write home about either" (Dennett, 2017, p. 278).

[19]Keats' phrase to define the creative mind, popularised by Bion and much espoused by dynamic psychotherapists (e.g., Symington and Symington, 1996).

[20]Russell's (2001) version of the Occam's razor principle of parsimony is: "Whenever possible, substitute constructions out of known entities for inferences to unknown entities." Thus, do we try to calibrate how new experience A is "like" known event B – and how it differs.

[21]In the mathematical analysis of non-linear systems, attractors are the set of numerical values toward which a system tends to evolve, from a wide variety of starting conditions. There is a possible link to the psychoanalytic notion of "fixation."

[22]For a detailed account of system/ego-centric, subjective values, and their role in transitions from proto to truly mental states as well as precision-weighed uncertainty representation, see Solms (2019).

[23]C.f., Einstein: "everything should be made as simple as possible, but not simpler" (Reader's Digest July 1977).

[24]The notorious "Milgram" (1974) experiments can be thought of in comparable terms. Those able to resist the seemingly sadistic urgings of the experimenter were using agency and top-down internal feedback–"I am under no obligation to continue with this."

prediction errors, driven by the need for a modicum of attachment security, however dysfunctional. An early task therefore in psychotherapy is to re-establish a degree of bio-behavioral synchrony. The patterns and rhythms of therapy help with this, as do the joint attention and affective mirroring (Holmes and Slade, 2017) typical of secure attachments. The more disturbed the individual, the longer this is likely to take–and it remains a fluctuating process varying from session to session and moment-to-moment within sessions.

Action is the prime means for improving the prediction and predictability of sensory sampling and thus minimizing prediction error. Clients suffering from depression are often in the thrall of cognitive errors that dominate their affective world: "everyone hates me," "I am useless," etc. These self-perpetuating–albeit parsimonious–priors not only bind free energy but also undermine agency and the ensuing accuracy of predictions. Passive helplessness pervades, interspersed with depressive auto-denigration. The "hand-holding" help of a therapist fosters action, initially in the form of verbal exploration. When things go well, depressive priors begin to be revised in the light of experience.

Bio-behavioral synchrony and the fostering of agency are probably common to all effective therapies. The remainder of our discussion focuses a free energy perspective on psychoanalytic therapies. Here, the role of "action" is less evident compared, say, with cognitive behavioral therapy (CBT), although the impulse to act–or "act out"–is an important focus for transferential and counter-transferential work. Indeed, choosing to seek help for psychological difficulties in itself implies a degree of agency. Furthermore, if conversation is seen in terms of "speech acts" analytic dialogue is in itself agency-enhancing.

DECOUPLING

We will touch on a number of key features of the analytic approach: free association, dreams, sexuality, reflective discourse, transference, and mentalising. All depend on "*decoupling*"–introducing a degree of "play" into the bottom-up/top-down surprise-minimizing articulations of everyday life (c.f., Holmes and Slade, 2017). In the presence of a modulating, moderating, affect-buffering therapist, surprise/energy unbound becomes tolerable and, when therapeutically scrutinized, extends the repertoire and range of a person's counterfactual realities, i.e., priors. Built into this model is both "creativity" and "destruction," in the sense that modification of error-prone priors entails their replacement with alternative hypotheses. The greater the range of prior hypotheses, the greater the opportunities for error-minimized binding and the less the need to resort to rigid, limited, or anachronistic priors, at the different levels of a hierarchy of generative models. This, in turn, enhances the adaptedness of the sufferer to their environment, including, *via* mentalising, the self. Part of the process makes the patient's model more accurate by revised belief-formation, and part by complexity reduction, especially in relation to resolution of conflict and trauma (Hopkins, 2016).

Decoupling From "Below": Free Association

Reducing prediction error is a complex multi-level and recursive process that reverberates up and down a series of interconnected message-passing hierarchies. "Bottom-up" does not refer simply to activity at sensory epithelia, but at each level of synaptic connection in a nested hierarchy of message-passing within canonical microcircuits throughout the brain. For example, Lanius et al. (2015) discuss decoupling between Prefrontal Cortex (PFC) and amygdala in post-traumatic states, and how, in the absence of top-down input from the PFC, patients attempt to dampen amygdala activity by resorting to substance abuse or self-harm. Observing these processes in a therapeutic setting forms a first step toward establishing reconnection and enhancing modulation of raw affect.

Barratt (2016) argues that Freud's greatest discovery, clinically and theoretically, was the concept and practice of "free association." Freud's (1916) image of this was that of the passenger in a train looking out of a window and observing the view as it flashes past. In free association, thoughts, interoceptive bodily sensations and effects, impulses, and images enter the mind "from below." As analysand and therapist collaboratively enter states of free-floating attention and negative capability, top-down constructions are temporarily set aside. Avoidant clients, with intellectual defenses, are both resistant to, and especially likely to benefit from joint attention to such free-associative experiences. With their co-regulatory sensitive period re-opened, they can explicitly attend to repressed feelings and fears. Free energy can now be minimized through prior modification and simulated action rather than repression. As in the study by Coan et al. (2006), the therapist's calming, containing, "slow-thinking" conversational presence generates "forms of feeling" (Mears, 2018), which the sufferer can discern and grasp rather than fearfully evade.

"Action Replay"

A crucial technique in the mentalising approach to psychotherapy with people suffering from Borderline Personality Disorder is a procedure known as "pressing the pause button" (Allen et al., 2008), when therapist and client interrupt the flow of their interactions in order to examine "what was going on between us just now." This disrupts automatic top-down/bottom-up pathways, making thoughts and behaviors available for scrutiny. An "event"–e.g., a client's sudden outburst of anger triggered by a therapist's holiday–may stimulate prolonged collaborative reflection, encompassing previous comparable interpersonal experiences. The client begins to tease out differences between a therapeutic "break" with a high probability of resumption, and a childhood history of being arbitrarily abandoned, leading to more complex and realistic posteriors about the reversibility of loss.

Dreams

During hours of dark, prediction errors inevitably increase. Applying the free energy model to the neurobiology of sleep, Hobson and Friston (2012) suggest that, when dreaming,

bottom-up sensory input and top-down prediction are de-coupled. In the absence of afferent input, potentially free-energetic–and so entropic–memory traces of the "days residue"[25] can be "bound" into parsimonious representations. *Via* synaptic pruning and consolidation of themes of affective significance, this "housekeeping" process reduces the chaotic complexity of everyday waking life.

Although this approach does not fully endorse the Freudian notion of dreams as disguised wish fulfillments, it sees dream themes as value-laden, replete with affective saliences which have not reached waking conscious awareness (c.f., Solms, 2013). At the same time, dreaming embodies *counter-factual simulation* or *virtual reality generation*. Triggered by the day's residue, possible future scenarios are played out in dreams helping to build a repertoire of free-energy minimizing priors, able to reduce prediction error when encountering future potentially traumatic events.[26] This process does not forestall emotional pain, but safeguards against, or at least postpones entropic surprise. Anything and everything is possible, thereby arming the dreamer against the unpredictability–improbabilities–of life.[27] Freud excluded undisguised trauma-related dreams from his wish-fulfillment theory. From a free energy perspective, dreaming reworks trauma so that it becomes "thinkable": "only a dream," or "that was then, inescapable, horrible; this is now, still painful, but tolerable" (Kinley and Reyno, 2017).

Transference

In the "flapping black object" example, an ambiguous stimulus presented itself to the subject, who saw something "*untoward*" out of the corner of his eye. Given the high degree of imprecision intrinsic to peripheral vision, this was interpreted in the light of plausible "prior" based on the previous day's experience–a possible poisoned bird. Disambiguation (Seth, 2015) followed: face-forward movement *toward* the object led to a revised "posterior." This illustrated how an interoceptive anxiety ("my nausea suggests that the bird could also have been poisoned by the weed spray") could shape an erroneous prior, leading to a maladaptive "perception," in which a picture of the world appropriate to the past (here the previous day) was carried over, or *transferred*, inappropriately, into the present.

According to Laplanche (2009, p. 93) "the analyst is the one who guards the enigma and provokes the transference." In his terms, the analyst is–like the world glimpsed in peripheral vision–an "enigmatic signifier," not perhaps an entirely "blank slate," but nevertheless embodying the reticence –creative ambiguity–inherent in analytic technique. Drawing on that ambiguity for therapeutic ends, the analyst receives and helps identify the patient's projected object relations or unconscious phantasies.[28]

From a free energy perspective, transference is an entrenched "prior," inaccessible to updating *via* active inference. In the classic Kleinian concept of "projective identification" (PI) (e.g., Ogden, 1992/2018), transference is jointly *enacted* by therapist and client. PI can be conceptualized in free energy terms as an attempt to *shape* the interpersonal world in the light of pre-existing phantasies, rather than to revise priors in the light of experience. For example, a therapist might "forget" to inform a PI-driven client about an upcoming break, having been induced unconsciously by the client's expectation of abandonment actually to do so.

But–exemplifying psychoanalysis' paradoxical capacity to snatch success from the jaws of defeat–such enactments also have the potential, as Winnicott (1974, p. 107) puts it, to "bring trauma within the arena of omnipotence" and hence be available for therapeutic work. The FEP point here is that one way to minimize surprise is actively to shape or seek out environments in accordance with one's priors, thereby eliminating the necessity to update them. Recognizing and exploring projective identification observes this process in action and offers a more flexible range of options for living out one's relationships. A crucial prerequisite is the therapist's countertransference awareness (Brenman Pick, 1985, 2018)–the capacity to be objective about one's own subjectivity.

Sexuality

The FEP is inherently temporal: *sensation* stimulates a prior, leading to *perception* and, *via* active inference, posterior *revision*. In the example, it was a "relief" to realize that the putative bird was a figment of imagination. In this FEP account, there is an affective arc of motivated tension, consummation, and resolution, in which the very binding of energy is rewarding. By deepening trust and discouraging premature closure of surprise, therapy fosters this expansion of the realm of desire.

Put another way–ambiguity and its resolution is both *exciting* and *rewarding*. To return to Laplanche (1987), enigma–which can be reformulated in this context as prediction error–is central to this process. In his neo-Oedipal model, the "breast" is a "sexual organ," but, for the naïve infant, one wrapped in mystery. The mother's loving sensuality in relation to her baby is suffused with a degree of eroticism which the child cannot fully comprehend.

Building on this, Target (2007) suggests that sexuality is the outstanding exception to the observation that joint attention and accurate affect-mirroring by caregivers underpins the development of the child's sense of self (Fonagy et al., 2002). In the realm of genital sexuality, parents typically distract, avoid, or punish rather than directly reflect the child's explorations and feelings of excitement. This, Target argues, leaves a residue of *mirroring-hunger*, whose resolution is postponed until sexual life begins in adolescence, and a suitable partner/other is found with whom a sexual duet for one can begin to be played. With its recurrent rhythms of desire and resolution, sexuality

[25]Freud's term.

[26]This account of dreaming can be compared to immunization in which overwhelming infection is prevented *via* prior exposure to attenuated forms of potential pathogens.

[27]Another parallel is with Bayesian weather forecasting. In the "numerical modeling method," the computer, "top-down," generates a large number of possible future weather patterns based on small differences in prior assumptions (Seth, 2014). Accuracy of priors is iteratively improved by posterior revisions which feed into the next day's forecast, and so on.

[28]E.g., Patient: "*have* you *got any children*?"; Analyst: "*that's a really interesting question–I wonder what has prompted it to come up today*?".

remains suffused with a continuing ambiance of enigma. Part of the mystery and paradox of sex is the tension between the fact that one can never fully "know" the other, and yet, through sex (genital and in any of the ontologically derived adult expressions of infantile, polymorphously perverse sexualities), one approaches their intimate being.[29] In FEP terms, sex "plays" with energy bound and unbound and their relationship to, among others, the reward system.

When sexuality permeates the analytic relationship as erotic transference, the "decoupling" virtual reality ambiance of psychoanalytic work, enables such feelings to be jointly mentalised, thereby enabling clients to develop a more explicit sense of the lineaments of their desires.

Therapeutic Conversations

While underpinned by pre-verbal bio-behavioral synchrony, psychotherapy is in essence a specialized form of conversation, or proto-conversation (Mears, 2018). Based on Strachey's (1934) classical paper on the "mutative interpretation," Lear (2011) suggests that change in psychoanalysis relies on the interplay between two conservational vectors. First is the mirroring and role-responsiveness as the analyst enters into patients' "idiolect," helping to delineate their unique way of seeing world and self-stamped vernacular, always trying to find the right words to capture the patient's "forms of feeling," without imposing her or his own emotional vocabulary. At this stage, from the patient's point of view, the top-down/bottom-up process runs smoothly, and, from a free energy perspective, un-"surprisingly."

But at some point, a discrepancy (or ambiguity) will inevitably arise, as the analyst fails to conform to the patient's top-down expectations. In the Strachey's 1930s account, the feared punitive father turns out to be benign; in a contemporary version, a patient's view of her analyst as abusive ("*you're just getting off on my misery; you don't really give a damn*") might be confounded by a degree of compassionate and committed concern. Conversely, patients' assumption that their therapists will be all-loving or all-forgiving comes up against confrontations, inflexible endings to sessions, the need to pay fees, etc. In the face of this discrepancy between desire and reality, patients do their best to maintain the *status quo*, clinging to past assumptions, attempting to evade the need to bind free energy with revised priors. This discrepancy then becomes the *point d'appui* of psychotherapeutic work.

From a free energy perspective, psychological ill health implies simplistic top-down models, and/or restricted sensory sampling, while structured complexity, as opposed to chaos or rigidity, is a mark of psychological health. Psychotherapy aims to increase the repertoire of its subjects' models of themselves and their environment. It is no mean task for analysts to challenge their patients, to break the mold of maladaptive energy binding, and to move psychic structures toward this augmented complexity. It is tempting to collude, "supportively" maintaining the *status quo*, or gratefully (if silently) accepting the drop-out of a "difficult" patient. Yet from a Bayesian perspective, Moutoussis et al. (2018) suggest that complexity is crucial to treatment success: too much, and there is no generalization from good therapeutic experiences to blighted everyday lives; but if complexity is simplistically minimized, this inhibits the risk-taking and "negative capability" needed for psychic change.

Recent research by Talia and his group (Talia et al., 2014, 2018) lends further experimental support to this model and to the attachment categories discussed earlier. Analyzing transcripts of psychotherapy sessions, they show how the nature of therapeutic dialogue depends on the attachment status of both client and therapist. Securely attached clients–and therapists–engage in turn-taking "duets," in which there is contact seeking, free exchange and modulation of affect and ideas. By contrast, insecurely attached people typically rebuff mutative speech acts. Their dialogue tends to be non-relational, with little affect-modulation, frequent backtracking, and repetitive interactive patterns.

The partial or occasionally total impasse created by these insecure speech patterns then becomes the focus of therapy. Painful affects–anxiety or misery–signal prediction errors, misalignment between wish and reality. But rather than leading to change, these become chronic and embedded. Psychotherapy mobilizes the active inference needed to resolve the impasse. The therapist enjoins the client to look at–mentalise–what is happening between them. Knowing that his or her hand is being metaphorically held, and that energy binding can be temporarily left to the therapist, the client can become more adventurous. In "duet for one" moments, initially fleetingly, therapist and client "sing" in ways that pertain to each and neither participant. Classical analytic geometry may encourage this–prone, in the absence of visual contact, patients take their analysts as part of themselves, drawing on the other's "priors"–i.e., verbal "interpretations"–to widen the range of available top-down models of the world and its possibilities.

CONCLUSION

In a perhaps slightly disingenuous moment of self-doubt, Friston (2010, p. 9) asks:

> "What does the free-energy principle portend for the future? If its main contribution is to integrate established theories, then the answer is probably 'not a lot'…[But it] could also provide new approaches to old problems that might call for a reappraisal of conventional notions."

Wiese (2015) argues that while FEP may in a Popperian sense be "unfalsifiable," it nevertheless represents a Kuhnian new paradigm. Our enthusiasm for the free energy model comes from the position of psychotherapy ecumenicalism (c.f., Holmes, 2002; Wampold, 2015; Holmes and Slade, 2017). We have argued that recovery from psychological ill-health

[29]FEP accounts for the impossibility of self-tickling (e.g., Hohwy, 2016) on the grounds that top-down priors thwart the necessary unexpectedness of a tickle. A similar argument could be mounted to explain the unsatisfactoriness of masturbation as opposed to relational sex.

is associated with enhancing the capacity to bind free energy and thereby facilitate prediction error minimization. Therapeutic procedures which foster these will be likely to be helpful, whatever their espoused brand name. These include the following: promoting agency; broadened sensory and interoceptive sampling, whether through CBT "experiments" or psychoanalytic free association; widening counter-factual simulation and the range of top-down hypotheses through dream-analysis and transference work; and fostering the capacity to modify priors in the light of experience, especially through the analysis of transference.

We have outlined some of the established interpersonal procedures which pave the way for these: bio-behavioral synchrony, epistemic trust, and turn-taking duet-for-one dialogue. From a research perspective, these features can be operationalized as benchmarks for assessing psychotherapy efficacy and procedural compliance. They help concentrate therapists and their supervisors' minds, and, we predict, improve clinical outcomes.

A final point in favor of the FEP is that it conceives psychotherapy, not as an esoteric concoction, but as a "natural kind," a specialized form of a general cultural phenomenon. Many aspects of cultural life–play, music, sport, drama, and iconography–depend on the top-down/bottom-up "decoupling" and mentalising which foster prediction error minimization,[30] and so enhance recovery and resilience.

[30]The actor-audience divide decouples meaning from action in a variant of Coan's hand-holding. Watching Shakespearean tragedy (Holmes, 2018)–or indeed a "horror movie"–extends the repertoire of top-down priors available for energy binding if and when real-life trauma strikes.

The homeostasis–psychological no less than physiological–essential, in Claude Bernard's (1974) famous phrase, to a free life, is vulnerable to the ever-present forces of entropy. The discrepancies between the affordances of the environment–which in our species' case is primarily interpersonal–and our inner models is the basis of prediction error, signaled by affective distress, leading, if unrevised, to entrenched mental pain or psychological illness. Learning to experience and resolve prediction error depends on the generative possibilities of intimate relationships. Where those fail or falter, psychotherapy provides a vital route to repair.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

TN is funded by a Wellcome Trust Principle Investigator Award and a NIH-NIDS award (5R01NS092701-03) to P. Read Montague.

ACKNOWLEDGMENTS

We thank Karl Friston, Jim Hopkins, Michael Moutoussis, Christoph Mathys, Julia Griem, Kristin White, and Barnaby Barrett for thoughtful comments on earlier drafts of this manuscript.

REFERENCES

Allen, B., Bendixsen, B., Fenerci, R. B., and Green, J. (2018). Assessing disorganized attachment representations: a systematic psychometric review and meta-analysis of the Manchester Child Attachment Story Task. *Attach. Hum. Dev.* 259–292. doi: 10.1080/14616734.2018.1429477

Allen, J., Fonagy, P., and Bateman, A. (2008). *Handbook of mentalizing in mental health practice*. (Arlington, VA: American Psychiatric Association Publishing).

Barratt, B. (2016). *Radical Psychoanalysis*. (London, England: Routledge).

Barrett, L. (2017). The theory of constructed emotion: an active inference account of interoception and categorisation. *Soc. Cogn. Affect. Neurosci.* 12, 1–23. doi: 10.1093/scan/nsw154

Bateman, A. W., and Fonagy, P. (2012). *Handbook of mentalizing in mental health practice*. (Arlington, TX: American Psychiatric Publishing).

Bernard, K., Meade, E., and Dozier, M. N. (2013). Parental synchrony and nurturance as targets in an attachment-based intervention: building on Mary Ainsworth's insights about mother-infant interaction. *Attach. Hum. Dev.* 15, 507–523. doi: 10.1080/14616734.2013.820920

Bion, W. (1962). *Learning from experience*. (London: Heinemann).

Bolis, D., and Schilbach, L. (2017). Beyond one Bayesian brain: modelling intra- and inter-personal processes during social interaction: commentary on "Mentalizing Homeostasis: the social origins of interoceptive inference" by A. Fotopoulou and M. Tsakiris. *Neuropsychoanalysis* 19, 35–38. doi: 10.1080/15294145.2017.1295215

Brenman Pick, I. (1985). Working through in the countertransference. *Int. J. Psychoanal.* 66, 157–166.

Brenman Pick, I. (2018). *Authenticity in the psychoanalytic encounter: the work of Irma Brenman Pick*. (Routledge).

Brown, H., Adams, R., Parees, I., Edwards, M., and Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cogn. Process.* 14, 411–427. doi: 10.1007/s10339-013-0571-3

Carhart-Harris, R. L., and Friston, K. J. (2010). The default mode, ego-functions and free energy: a neurobiological account of Freud's idea. *Brain* 133, 1265–1283. doi: 10.1093/brain/awq010

Carroll, L. (1871/2009). *Through the looking glass and what Alice found there*. (Westport Eire: Everson).

Cittern, D., Nolte, T., Friston, K., and Edalat, A. (2018). Intrinsic and extrinsic motivators of attachment under active inference. *PLoS One* 13:e0193955. doi: 10.1371/journal.pone.0193955

Coan, J. (2016). "Attachment and neuroscience" in *Handbook of attachment. 3rd Edn.* eds. J. Cassidy and P. Shaver (New York, NY: Guilford Press), 242–269.

Coan, J. A., Schaefer, H. S., and Davidson, R. J. (2006). Lending a hand: social regulation of the neural response to threat. *Psychol. Sci.* 17, 1032–1039.

Conant, R. C., and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 89–97.

Debbané, M., and Nolte, T. (2019, forthcoming). "The neurobiology of mentalising" in *Handbook of mentalizing in mental health practice*. 2nd Edn. eds. P. Fonagy and A. Bateman (in press).

Dennett, D. (2017). *From bacteria to bach and back*. (London: Allen Lane).

Feldman, R. (2015a). Sensitive periods in human social development: new insights from research on oxytocin, synchrony, and high-risk parenting. *Dev. Psychopathol.* 27, 369–395.

Feldman, R. (2015b). The adaptive human parental brain: implications for children's social development. *Trends Neurosci.* 38, 387–399.

Fonagy, P., and Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy* 51, 372–380.

Fonagy, P., Gergely, G., Jurist, E., and Target, M. (2002). *Affect regulation, mentalization, and the development of the self*. (New York, NY: Other Press).

Fotopoulou, A., and Tsakiris, M. (2017). Mentalizing homeostasis: the social origins of interoceptive inference. *Neuropsychoanalysis* 19, 3–28. doi: 10.1080/15294145.2017.1294031

Freud, S. (1895/1950). *Project for a scientific psychology. SE 1 95-397*. (London: Hogarth).

Freud, S. (1916). *Introductory lectures in psychoanalysis. SE16 17*.

Freud, S. (1925). *An autobiographical study SE 20 p3*. (London: Hogarth).

Friston, K. (2010). The free energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475.

Friston, K. (2018). Am I self-conscious (or does self-organisation entail self-consciousness?). *Front. Psychol.* 9. doi: 10.3389/fpsyg.2018.00579

Friston, K., and Frith, C. (2015). A duet for one. *Conscious. Cogn.* 36, 390–405. doi: 10.1016/j.cogn.2014.12.003

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Comput.* 29, 2633–2683. doi: 10.1162/neco_a_00999

Frith, C. (2012). The role of metacognition in human social interactions. *Philos. Trans. R. Soc. B.* 367, 2213–2223. doi: 10.1098/rstb2012.0123

Gershman, S. J. (2017). Predicting the past, remembering the future. *Curr. Opin. Behav. Sci.* 17, 7–13. doi: 10.1016/j.cobeha.2017.05.025

Gibson, E., and Walk, R. (1960). Visual cliff. *Sci. Am.* 202, 64–71. doi: 10.1038/scientificamerican0460-64

Gibson, J. (1986). *The ecological approach to visual perception*. (Hillsdale, New Jersey: Lawrence Erlbaum Associates).

Hamilton, A. F., and Lind, F. (2016). Audience effects: what can they tell us about social neuroscience, theory of mind and autism? *Cult. Brain* 4, 159–177. doi: 10.1007/s40167-016-0044-5

Hobson, J., and Friston, K. (2012). Waking and dreaming consciousness: neurobiological and functional considerations. *Prog. Neurobiol.* 98, 82–98. doi: 10.1016/j.pneurobio.2012.05.003

Hofer, M. (2002). Clinical implications drawn from the new biology of attachment. *J. Infant Child Adolesc. Psychother.* 2, 157–162. doi: 10.1080/15289168.2002.10486425

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

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Holler, J., Kendrick, K., Casillas, M., and Levinson, S. D. (2015). Turn-taking in human communicative interaction. *Front. Psychol.* doi: 10.3389/fpsyg.2015.01919

Holmes, J. (2002). *The search for the secure base*. (London: Routledge).

Holmes, J. (2010). *Exploring in security: Towards an attachment-informed psychotherapy*. (London: Routledge).

Holmes, J., and Slade, A. S. (2017). *Attachment in therapeutic practice*. (London: SAGE).

Holmes, J. (2018). Perdita and Oedipus: a tale of two adoptions. *Br. J. Psychother*. (in press).

Hopkins, J. (2016). Free energy and virtual reality in neuroscience and psychoanalysis: a complexity theory of dreaming and mental disturbance. *Front. Psychol.* 7:922. doi: 10.3389/fpsyg.2016.00922

Isomura, T., and Friston, K. (2018). In vitro neural networks minimise variational free energy. *Scientific reports.* 8:16926. doi: 10.1038/s41598-018-35221-w

Kahneman, J. (2011). *Thinking: Fast and slow*. (London, England: Allan Lane).

Kinley, J., and Reyno, S. (2017). Advancing Freud's dream: a dynamic-relational neurobiologically informed approach to psychotherapy. *Neuropsychoanalysis*. doi: 10.1080/15294145.2017.1367260

Kirchhoff, M. (2017). Predictive brains and embodied enactive cognition: an introduction to special issue. *Synthese* 195, 2355–2366.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface*. 15:20170792. doi: 10.1098/rsif.2017.0792

Klein, M. (1946). Notes on some schizoid mechanisms. *Int. J. Psychoanal.* 27:99.

Klein, M. (1997). *Envy and gratitude: And other works, 1946-1963*. (London: Random House).

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719.

Knox, J. (2010). *Self-agency in psychotherapy*. (New York, NY: Norton).

Lakoff, S., and Johnson, M. (2003). *The metaphors we live by. 2nd Edn.* (Chicago, IL: University of Chicago Press.)

Lanius, R. A., Frewen, P. A., Tursich, M., Jetly, R., and McKinnon, M. C. (2015). Restoring large-scale brain networks in PTSD and related disorders: A proposal for neuroscientifically-informed treatment interventions. *European Journal of Psychotraumatology* 6.

Laplanche, J. (1987). *New foundations for psychoanalysis*. Translated by D. Macey. (Oxford: Blackwell).

Laplanche, J. (2009). "Transference: its provocation by the analyst" in *Reading French psychoanalysis*. eds. J. Birkstead-Breen, S. Flanders and A. Gibeault. Translated by J. Cheshire. (London: Routledge).

Lear, J. (2011). *A case for irony*. (Cambridge, MA: Harvard University Press.)

Leichsenring, F. (2008). Effectiveness of long-term psychodynamic psychotherapy. *JAMA* 300, 1551–1565.

Leichsenring, F., Luyten, P., Hilsenroth, M. J., Abbass, A., Barber, J. P., Keefe, J. R., et al. (2015). Psychodynamic therapy meets evidence-based medicine: a systematic review using updated criteria. *Lancet Psychiatry* 2, 648–660. doi: 10.1016/S2215-0366(15)00155-8

Leonidaki, V., Lemma, A., and Hobbs, I. (2018). The active ingredients of dynamic interpersonal therapy (DIT): an exploration of client's experiences. *Psychoanal. Psychother.* 32, 140–156. doi: 10.1080/02668734.2017.1418761

Maier, S. F., and Seligman, M. E. (2016). Learned helplessness at fifty: insights from neuroscience. *Psychol. Rev.* 123, 349–367. doi: 10.1037/rev0000033

Mears, R. (2018). *The poet's voice in the making of mind*. (London: Routledge.)

Mellor, M. (2018). Making worlds in a waking dream: where Bion intersects Friston on the shaping and breaking of psychic reality. *Front. Psychol.* (in press). 9. doi: 10.3389/fpsyg.2018.01674

Milgram, S. D. (1974). *Obedience to authority*. (New York: Harper & Row).

Moutoussis, M., Shahar, N., Hauser, T. U., and Dolan, R. J. (2018). Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Comput. Psychiatry* 2, 50–73. doi: 10.1162/CPSY_a_00014

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R. J., and Friston, K. J. (2014). A formal model of interpersonal inference. *Front. Hum. Neurosci.* 8:160. doi: 10.3389/fnhum.2014.00160

Nolte, T., Bolling, D. Z., Hudac, C., Fonagy, P., Mayes, L. C., and Pelphrey, K. A. (2013). Brain mechanisms underlying the impact of attachment-related stress on social cognition. *Front. Hum. Neurosci.* 7:816. doi: 10.3389/fnhum.2013.00816

Nolte, T., Campbell, C., and Fonagy, P. (2019). "Social communicative processes in severe personality disorder" in *The psychotherapy-neurobiology-pharmacology intervention triangle*. eds. V. Bizzari, J. Gonçalves and J. G. Pereira (New York: Vernon Press).

O'Keefe, J. (1978). *The hippocampus as a cognitive map*. ISBN 978-0198572060.

Ogden, T. (1992/2018). *Projective identification and psychoanalytic technique*. (London: Routledge.)

Ogden, T. (1994). The analytic third: working with intersubjective clinical facts. *Int. Psychopathol.* 27, 369–395.

Palmer, C. J., Seth, A. K., and Hohwy, J. (2015). The felt presence of other minds: predictive processing, counterfactual predictions, and mentalising in autism. *Conscious Cogn.* 36, 376–389. doi: 10.1016/j.concog.2015.04.007

Powers, A. R. 3rd, Bien, C., and Corlett, P. R. (2018). Aligning Computational Psychiatry With the Hearing Voices Movement: Hearing Their Voices. *JAMA Psychiatry.*

Ridley, M. (1993). *The red queen: Sex and the evolution of human nature*. (London: Penguin.)

Rudrauf, D. (2014). Structure-function relationships behind the phenomenon of cognitive resilience in neurology: insights for neuroscience and medicine. *Adv. Neurosci.* 2014, 1–28. doi: 10.1155/2014/462765

Rudrauf, D., and Debbané, M. (2018). Building a cybernetic model of psychopathology: beyond the metaphor. *Psychol. Inq.* 29, 156–164. doi: 10.1080/1047840X.2018.1513685

Russell, B. (2001). *The collected papers of Bertrand Russell, Volume 9: Essays on language, mind and matter: 1919–1926*. J. G. Slatered. (London and New York: Russel), 160–179.

Schröedinger, E. (1944). *What is life? The physical aspect of the living cell*. (Cambridge: Cambridge University Press).

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. *Cogn. Neurosci.* 5, 97–118.

Seth, A. K. (2015). "Inference to the Best Prediction" in *Open MIND*. eds. T. K. Metzinger and J. M. Windt (Frankfurt am Main: MIND Group).

Seth, A. K., and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philos. Trans. R. Soc. B* 371:20160007. doi: 10.1098/rstb.2016.0007

Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. *Am. Psychol.* 65, 98–109. doi: 10.1037/a0018378

Solms, M. (2013). The conscious id. *Neuropsychoanalysis* 15, 5–19. doi: 10.1080/15294145.2013.10773711

Solms, M. (2015). *The feeling brain: Selected papers on neuropsychoanalysis*. (London: Routledge).

Solms, M. (2019). The hard problem of consciousness and the free energy principle. *Front. Psychol.* 9. doi: 10.3389/fpsyg.2018.02714

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiol. Behav.* 106, 5–15.

Strachey, J. (1934). The nature of the therapeutic action in psychoanalysis. *Int. J. Psychoanal.* 15, 126–159.

Symington, N., and Symington, J. (1996). *The clinical thinking of Wilfrid Bion*. (London: Karnac.)

Talia, A., Daniel, S. I., Miller-Bottome, M., Brambilla, D., Miccoli, D., Safran, J. D., et al. (2014). AAI predicts patients' in-session interpersonal behavior and discourse: a "move to the level of the relation" for attachment-informed psychotherapy research. *Attach. Hum. Dev.* 16, 192–209. doi: 10.1080/14616734.2013.859161

Talia, A., Muzi, L., Lingiardi, V., and Taubner, S. (2018). How to be a secure base: therapists' attachment representations and their link to attunement in psychotherapy. *Attachment & human development* 1–18.

Target, M. (2007). Is our sexuality our own? An attachment model of sexuality based on early affect mirroring. *Br. J. Psychother.* 23, 517–530. doi: 10.1111/j.1752-0118.2007.00048.x

Taylor, D. (2015). Pragmatic randomized controlled trial of long-term psychoanalytic psychotherapy for treatment resistant depression: the Tavistock Adult Depression Study (TADS). *World Psychiatry* 14, 312–321.

Tervo, D. G., Tenenbaum, J. B., and Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Curr. Opin. Neurobiol.* 37, 99–105. doi: 10.1016/j.conb.2016.01.014

Tottenham, N. (2014). "The importance of early experiences for neuro-affective development" in *The neurobiology of childhood*. eds. S. Anderson and D. Pine, vol. 16. (Berlin: Springer), 109–129.

Wampold, B. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry* 14, 270–277. doi: 10.1002/wps.20238

Wiese, W. (2015). "Perceptual presence in the Kuhnian-Popperian Bayesian Brain" in *Open MIND*. eds. T. Metzinger and J. M. Windt: 35(C).

Wilson, D. (2002). *Darwin's cathedral*. (Chicago, IL: University of Chicago Press).

Winnicott, D. (1974). Fear of breakdown. *Int. Rev. Psychoanal.* 1, 103–107.

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.

The handling editor declared a shared affiliation, though no other collaboration, with one of the authors TN.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 28 August 2019
doi: 10.3389/fpsyg.2019.01983

Psychoanalysis and Neuroscience: The Bridge Between Mind and Brain

Filippo Cieri[1]* *and Roberto Esposito*[2]

[1]Department of Neurology, Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV, United States, [2]Department of Radiology, Azienda Ospedaliera Ospedali Riuniti Marche Nord, Pesaro, Italy

In 1895 in the Project for a Scientific Psychology, Freud tried to integrate psychology and neurology in order to develop a neuroscientific psychology. Since 1880, Freud made no distinction between psychology and physiology. His papers from the end of the 1880s to 1890 were very clear on this scientific overlap: as with many of his contemporaries, Freud thought about psychology essentially as the physiology of the brain. Years later he had to surrender, realizing a technological delay, not capable of pursuing its ambitious aim, and until that moment psychoanalysis would have to use its more suitable clinical method. Also, he seemed skeptical about phrenology drift, typical of that time, in which any psychological function needed to be located in its neuroanatomical area. He could not see the progresses of neuroscience and its fruitful dialogue with psychoanalysis, which occurred also thanks to the improvements in the field of neuroimaging, which has made possible a remarkable advance in the knowledge of the mind-brain system and a better observation of the psychoanalytical theories. After years of investigations, deriving from research and clinical work of the last century, the discovery of neural networks, together with the free energy principle, we are observing under a new light psychodynamic neuroscience in its exploration of the mind-brain system. In this manuscript, we summarize the important developments of psychodynamic neuroscience, with particular regard to the free energy principle, the resting state networks, especially the Default Mode Network in its link with the Self, emphasizing our view of a bridge between psychoanalysis and neuroscience. Finally, we suggest a discussion by approaching the concept of Alpha Function, proposed by the psychoanalyst Wilfred Ruprecht Bion, continuing the association with neuroscience.

Keywords: psychoanalysis, neuroscience, free energy principle, resting state network, default mode network

> *The real difference lies rather in the fact that the kind and direction of the physical vectors in Aristotelian dynamics are completely determined in advance by the nature of the object concerned. In modern physics, on the contrary, the existence of a physical vector always depends upon the mutual relations of several physical facts, especially upon the relation of the object to its environment.*
>
> Levin (1935), p. 35.

OPEN ACCESS

Edited by:
Jim Hopkins,
University College London,
United Kingdom

Reviewed by:
Patrick Connolly,
Hong Kong Shue Yan University,
Hong Kong
Karl Friston,
University College London,
United Kingdom

***Correspondence:**
Filippo Cieri
filippocieri@gmail.com

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *10 April 2019*
Accepted: *13 August 2019*
Published: *28 August 2019*

Citation:
Cieri F and Esposito R (2019) Psychoanalysis and Neuroscience: The Bridge Between Mind and Brain. Front. Psychol. 10:1983. doi: 10.3389/fpsyg.2019.01983

INTRODUCTION

Cognitive neuroscience has made remarkable advances also thanks to the progresses in neuroimaging techniques, such as Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI). One of the most important aims of this discipline is the understanding of human brain function. The dialogue between cognitive neurosciences and psychoanalysis is not new, but recently it has become more prolific in the exploration of the relationship between mind and brain, already wished for by Freud more than a century ago and, among others, by the Nobel Prize winner Kandel (1999), when he asserts that psychoanalysis still represents the most coherent and intellectually satisfying view of the mind and can help neurobiologists to plan their work.

To date neurosciences do not provide a consistent, consensual and comprehensive theory about the human brain-mind function, however it is a paramount tool in order to investigate structures and functions about mind-brain in its physiological and pathological development. Psychoanalysis flourished more than a century ago, but despite the first enthusiasm derived from initial fruitful dialogue with neuroscience, we are rather far from understanding the biological basis for all psychoanalytic theoretical frameworks, and this should not be the common goal of psychoanalysts or neuroscientists in their daily work. Although neuroscience and psychoanalysis share the same scientific object of interest, meant as a knowledge in-depth analysis about the functioning of mind-brain system, they use different tools of investigation, different methods and different languages, which requires a separation and distinction, albeit within an ongoing and steady dialogue between the two fields.

Since the birth of psychoanalysis, Freud has attempted to maintain a focus on the neurophysiological phenomena underlying the psychic processes observed. He had to abdicate the pursuit of his dream, first because the technologies available at his time were not sufficiently advanced to seek his neuroscientific ambition and on the other side because of his skepticism about the widespread phrenologic view and the disposition to fit any mental process in its specific brain region. This typical localizationist view was back in vogue after the important Paul Broca's discoveries in 1861, about the areas of language and his homonymous aphasia, determined by the lesion of an area that still maintains today Broca's name. Interest was renewed but not completely new, given that since the beginning of the 19th century Franz Joseph Gall, pioneer in the study of the cerebral cortex, focused his neuroanatomical investigation on the attribution of specific psychic functions to specific brain structures. In the first phrenology view Gall believed that man's moral and intellectual faculties were innate and strictly connected to the organization of the brain. He also proposed a localizationism view of the brain where single regions were responsible for a given mental faculty, and he finally suggested that the development of mental faculties in an individual would lead to a growth or larger development in the sub-region responsible for them.

The phrenology with Gall and the Broca's localizationism were both a kind of view, an attitude which never satisfied Freud, skeptical about the possibility of embedding every single mental function in its own presumed brain region, frustrated by a static and essentially mechanistic approach, immediately aware about the simplistic, reductive and reductionist imprint of the method. On the contrary he was starting to develop an increasingly dynamic vision of mind and brain.

The old reduction of mental functions to brain structures still finds today numerous supporters and attempts. As Tretter and Löffler-Stastka (2018) pointed out, this attempt encompasses a lot of well-known epistemological, methodological, and conceptual inconsistencies (Block, 1980; Chalmers, 1996; Craver, 2007).

After his experience at the Salpêtriere Hospital in Paris, Freud began to think about large brain networks with a variety of functions, with mutual activation and inhibition properties. An inference that anticipated the concept of *neural networks*, as large brain areas able to activate and inhibit, depending on the activity performed. During the same period, he matured the idea that in the brain there were no isolated centers, or autonomous functions, but instead systems responsible for complex cognitive purposes, composed by several regions, able to be modified by experience.

This dialectic reflects modern day distinctions between functional segregation (or specialization) and integration that have dominated thinking about modern brain imaging. In other words, does one understand distributed processing in terms of specialized regions or the integration and coordination of neuronal activity across brain hierarchies?

Freud's concept of large brain networks – against the localizationism and reductionist view – showed to anticipate a road that would lead to the concept of complex functional systems, developed more than 50 years later by Lurija (1976), founder of neuropsychology, whose central aim was to reject the idea of reductionism in psychology.

In his Project (Freud, 1895/1963, 1895/1966), Freud tried to explain at various levels the physiological basis of memory, hypothesizing that one of the neurophysiological prerequisites necessary for this function was a system of barriers, which he named "contact-barrier." He used this term to describe the neurophysiological entity which 11 years later, Charles Scott Sherrington named synapses.

The anticipation of wide and widespread systems dedicated to the realization of cognitive purposes, which today we know as neural networks, becomes impressive within the parallelism between the functions of the ego and specific neural networks, particularly Default Mode Network (DMN), one of the most studied brain networks by the neuroscientific community. Raichle and colleagues coined the term "Default Mode" in 2001 (Raichle et al., 2001); they used PET and described a specific brain state of "rest," a concept intended to quickly become fundamental in the study of the brain. DMN's functions seem to play the same mediation function attributed by Freud to the ego (Carhart-Harris and Friston, 2010). In particular, within the DMN, specific regions seem to support the monitoring phases regarding psychological state (Phan et al., 2002), considered areas in which the internal stimuli (bodily and proprioceptive sensations) and inputs from the external

environment (e.g., visual and auditory) converge for their integration and development. We will discuss this specific aspect further later.

In the last 10 years, the free energy principle has become *the royal road* in the dialogue between neuroscience and psychoanalysis, *the bridge* between mind and brain. It is linked to the work of Friston and colleagues (Friston et al., 2006; Friston, 2010) and it describes the function of the mind-brain system as any other adaptive biological system, connecting psychological sciences, neurosciences and related fields in perfect confluence and synergy with psychoanalytic concepts (Hopkins, 2012). This approach shows many similarities with typical psychoanalytical concepts as the secondary principle of mental functioning (Carhart-Harris and Friston, 2010), unconsciousness and motivation (Hopkins, 2012), complexity of emotions in attachment (Hopkins, 2015), wish fulfillment within dreaming (Hopkins, 2016), quantitative approach for a formulation of conflict (Hopkins, 2016), and the energic theory within psychoanalysis (Connolly, 2016).

The free energy principle considers the brain as a hierarchical, inferential, Helmholtzian machine, where large-scale intrinsic networks occupy supraordinate levels of hierarchical brain systems that try to optimize their representation of the sensorium, minimizing the amount of free energy (Friston et al., 2006). It represents a process formally close to Freudian metapsychology, in which Freud distinguished two ways of mental functioning: the primary and the secondary processes, corresponding to the pleasure and reality principle, respectively. According to the free energy principle the Bayesian brain uses the Bayesian probability approach to formulate perception as a constructive process based on internal or generative models (Knill and Pouget, 2004; Friston et al., 2006). The brain, with its personal model of the world (von Helmholtz, 1962; Gregory, 1980), tries to optimize this model using new information coming from sensory inputs (Ballard et al., 1983; Friston, 2005). These Bayesian formulations represent a fundamental advance over earlier formulations of optimization in the brain that inherit from behaviorism (e.g., reinforcement learning and optimal control) by explicitly considering (Bayesian) beliefs. In other words, the imperatives for neuronal message passing are framed in terms of belief updating. In this setting, the free energy functional that underwrites active inference under the FEP is actually a functional (i.e. function of a function) of probabilistic beliefs. This is important because it furnishes a calculus of beliefs that is much easier to relate to psychoanalytic constructs (relative to cost or value functions used in behaviorism).

At Freud's time, one of the most important mental disorders was hysteria, quite widespread among the population at the end of the 1800s. Within this syndrome, neurologists, psychiatrists, psychologists and psychoanalysts – mostly under the debate of the schools of Salpetrière and Nancy and their leaders Charcot and Bernheim, respectively – were especially interested about the link between mind and body. Among these scholars there was a young Freud as well, who tended to attribute a central role to the body in its connection to the mind. In his Studies on Hysteria (Freud, 1895/1963, 1895/1966), he observed somatic symptoms associated with mental disorders, underlying a close *psychosomatic* connection, after which he elaborated the concept of drive (Instincts and their Vicissitudes, Freud, 1915). With his second topical, in the ego and the id, the psychoanalytic meaning of the body assumes even greater centrality: "the ego is first and foremost bodily entity" (Freud, 1923). Freud thought of the ego as an entity derived from bodily sensations, especially from the sensations coming from the surface of the body. Over the years and deepening of clinical practice, psychoanalysis has begun to configure the link between body and mind not only as fundamental in structuring the ego and with a key role in the relationship with reality, but also with a vision of greater continuity and dynamic fluidity between organic and psychic dimensions, in which the free energy principle represents a useful bridge for the comprehension and communication between neuroscience and psychoanalysis.

Bion (1959) elaborated Freud's writing "Formulations on the Two Principles of Mental Functioning" (Freud, 1911), particularly focusing his observation on the body and sensory organs as instruments of access to the perception of reality. Bion considered thought and emotion as inseparable components, underlying the central role of the body as the start for the thought phenomena. He focused his observation on sense organs as instruments of access to the perception of reality, explaining how thought is a direct evolution of body sensations. Bion reversed the traditional philosophical conception in which mind produces thoughts: in his theory, there are first thoughts, and mind arises to think them. In other words, mind-body unit is constituted by the body that is in contact with external reality, then there are internal and external sensations, their perception and elaboration that generate emotions, moods and finally thoughts that we eventually perceive as products of the mind (Ciocca, 2015). All this process is supported by α-function. The capacity to transform the sense impressions related to an emotional experience, into α-elements is described as continuous in both sleeping and waking states (Mellor, 2018).

Bion builds a unique model of mind functioning, where the mind faces up continuously to new experiences that cause an emotional impact (positive or negative), and he proposes a general model of functioning of mind in which mental growth depends on the ability of the mind to digest new experiences[1].

In our manuscript, we try to move through development of a systemic view of the mind, taking in considerations of psychoanalytic models from Freud to Bion, their connections with modern neuroscience and neural networks, underlying the role of the Free Energy Principle as a bridge between mind and brain. We try to assume a methodological parallelism as it was seen for instance by the founder of General Systems Theory, Ludwig von Bertalanffy (von Bertalanffy, 1967), in which a systemic non-reductive multi-level approach might offer better options for integration (Miller, 1978; Tretter and Löffler-Stastka, 2018).

[1]Technically, the digestion of new experiences corresponds to the "data assimilation" or "evidence accumulation" implicit in "belief updating" under the FEP.

PSYCHOANALYSIS AND FREE ENERGY PRINCIPLE

Despite the extraordinary progresses made by psychology and cognitive neurosciences – among others the deepening of memory functioning and neuroimaging methods – there are few global theories regarding the operating of mind-brain, and no generalized or complete agreement in the neuroscientific community, not even with regard to the meaning of consciousness and unconscious. The proposal of the free energy principle (FEP henceforth) for adaptive systems provides a unified theory of action, perception, and learning (Friston, 2009).

According to Friston (2009), the FEP argues that any self-organizing system in nonequilibrium steady-state with its environment must minimize its free energy, describing how adaptive systems (as biological organisms) resist a natural tendency to disorder (Ashby, 1947; Kauffman, 1993; Friston, 2009). The defining characteristic of biological systems is their attempt to maintain a state of balance toward the constant changes in the environment (Ashby, 1947; Kauffman, 1993), as any homeostatic principle. In the allostatic principle proposed by Sterling and Eyer in 1988, also called a major revision (McEwen, 2004), replacement (Sterling, 2004) of the classical theory of homeostasis, the brain is identified as the central mediator of ongoing system-wide physiological adjustment to environmental challenge (Sterling and Eyer, 1988; McEwen, 2007). Both homeostasis and allostasis are endogenous systems engaged in maintaining an internal balance of the organism, coping with the continuous internal and external changes.

FEP rests on the idea that all biological systems instantiate a hierarchical generative model of the world that implicitly minimizes its sensory entropy by minimizing the level of its free energy (Ramstead et al., 2018). In other words, self-organizing systems, including human being as an example of biological organism, must resist the distributed effects of a natural increase in entropy for their existence, development, and evolution by trying to minimize free energy.

According to Friston et al. (2015a), these self-organizing systems must have a specific identifiable boundary condition: the so-called Markov blanket, which acts as a protective screen, described by Friston et al. (2015a) as a veil through which we are able to recognize and distinguish an internal side from an external environment of an organism, inferring the external or internal causes of sensations, perceptions, or changes. The Markov blanket is not only a protective screen thanks to which we can infer the external causes of the sensorium, it is also operates as a "projection screen" onto which sensory impressions are cast – that are actively solicited by habitual mechanisms (i.e., reflexes mediated by active states), which are used to make sense of the world (Friston et al., 2015a). As any other screen, the Markov blanket allows the separation of an internal dimension from an external environment of an organism, as the case of the cell, which typically represents an immediate and primordial example of a living system with a Markov blanket (Kirchhoff et al., 2018; Mellor, 2018). In this view, the boundaries of a neuron are defined by the external cell membrane, called *plasmalemma*, which is the Markov blanket of the neuron, ensuring separation and identification between an external environment and an internal state, protecting the cell from the external environment, guaranteeing its functions also through this distinction of environments and different electrical charges.

As Connolly (2018) points out, the Freudian energic theory has been widely critiqued by some authors, such as the lack of empirical evidence from neuroscience of the energic processes as described in "The Project" (Zepf, 2010), or the well-known critique by Rapaport (1960), which underlined the impossibility of direct energic processes observation in the clinical situation (Connolly, 2018). Although we cannot observe directly the energy and measure it during the clinical situation (the usefulness of which would be rather limited in any case), we can see the implicit or explicit physiological and psychological attempts by the patients to avoid surprises, especially with non-psychotic patients, who often seem to "prefer and choose" unpleasant and/or painful states (e.g., repetitive, anxious, depressive), but perfectly known, compared to a choice of a change, which apparently could bring emotional, personal, social, or professional benefits, but includes an unavoidable change, surprise, novelty, a new unknown and therefore strongly aversive state. These attempts are definitely psychological and physiological, thus of physical nature.

The mind-brain system tries to maintain the states within physiological bounds, which means trying to maintain a condition where the chances of surprise are minimized, ensuring that internal states remain within physiological and acceptable bounds for the organism. These kinds of attempts are often steady and strenuous, and they are felt as real imperatives in clinical, psychotherapeutic/psychoanalytic settings, in which patients try to avoid surprises and novelties often might be represented by a change of job, partner, or change of any other current distressing situation. A clinical frame experienced as painful by a patient who feels stuck but somehow safer in the current situation experienced as suffering but known, and for this reason is "preferred" to a new one that is potentially and surprisingly dangerous. During this clinical moment it is possible to observe the individual's effort engaged in his *data assimilation* from the environment, comparing it with the internal data and reality, trying to keep down the entropy levels and minimizing the possibility of surprise, avoiding excessive energy investment in an extremely hard and tiring psychophysical work.

As we mentioned, in the Bayesian brain principle the brain acts having a model of the world (von Helmholtz, 1962; Gregory, 1980), working through active inference, as an inference machine, generating actively predictions (von Helmholtz, 1962; Gregory, 1980; Dayan et al., 1995; Friston, 2010), and with this principle the brain tries to resist to a natural tendency to disorder, maintaining a sustained and homoeostatic exchange with its environment. According to Friston (2010), the brain's attempt to minimize the variations of free energy (maximizing Bayesian model evidence) not only provides a principled explanation for perceptual (Bayesian) inference in the brain but can also explain action and behavior (Ortega and Braun, 2010). Helmholtz's model about the brain

as an inference machine (Helmholtz, 1866/1962; Dayan et al., 1995) remains a key concept in neurobiology (Gregory, 1980) and psychology.

In this framework the brain works continuously trying to find pattern – thereby reducing free energy and minimizing surprise[2] – an effort that tries to reduce the free energy, minimizing the surprise from the system. This effort for the most part takes place in a completely implicit, unconscious way. For the individual, surprise means high level of free energy, leading the system to possible incorrect, erroneous, and unreliable predictions in relation to the world around it (Friston et al., 2015b). In psychoanalysis, this inaccurate prediction is translated as a poor testing of reality. Optimal reality testing would therefore require a minimization or reduction in free energy (surprise). This is implicit in belief updating that converts prior beliefs into posterior beliefs that minimize free energy. This can be thought of as the mathematical image of "binding energy" in a Freudian sense, where this "binding" occurs within the boundary established by the Markov blanket (Friston et al., 2015b).

The link between free energy and complexity is straightforward: free energy or surprise can be decomposed into complexity minus accuracy. This means that minimizing surprise (or maximizing model evidence) entails a maximization of accuracy in terms of explaining sensory impressions while, at the same time, minimizing complexity. This corresponds to Occam's principle and says that we try to find the simplest possible explanations that provide an accurate account of our sensorium.

According to Hopkins (2016), the FEP allows one to observe how the statistical conception of complexity employed by Friston and colleagues relates to emotional conflict and trauma; how symptoms as well as dreams can be understood in terms of complexity-reduction; how in a similar way REM dreaming reduces complexity though the consolidation/reconsolidation of memory; and how complexity and the mechanisms that have evolved to reduce it seem to play a key role for the understanding of mental disorders.

FEP today has a fundamental role in the dialogue with neurosciences and within psychoanalysis itself, describing an important model in understanding and deepening the functioning of the mind-brain system, offering a bridge between neural and psychological processes. As pointed out by Hopkins (2016), this linking of complexity, dreaming, and disorder also indicates that Freud and free association offer a clear and sharp path with cognitive science, free energy neuroscience, and computational psychiatry in order to create a consistent and solid connection between the psychological and neuroscientific views (Hopkins, 2016).

[2]An interesting corollary of surprise minimization is that we are compelled to seek out novelty, because novelty affords the opportunity to reduce expected surprise or uncertainty. One common feature found in non-psychotic patients concerns a certain "extension", or shift of discomfort from surprise to novelty in general, as if any novelty could lead to a risk of compromising the system. This mechanism is easily found in depressive, anxious or obsessive patients, in which we can observe how they try to avoid and defend themselves from novelty, as well as from surprise, repeating their same known and "safe" patterns.

Thanks to the dialogue with neuroscience and the FEP, Freud's free energy can be related to the potentially unifying paradigm advanced by Friston and colleagues, giving us the opportunity to better understand the mind-brain system in functional and dysfunctional disposition, through the investigation of psychoanalytic theory and models.

RESTING STATE NETWORKS AND THE DEFAULT SELF

In the Ego and the Id (Freud, 1923), Freud claims that the ego is not master in its own house, in other words the conscious instance is neither the only responsible nor the most important factor for the human behavior. The ego is influenced by the contradictory impulses of other instances, whose actions are often hidden. These other instances are the id, present at birth, established by constitution, consisting of impulses and instincts that originate from the bodily organization, finding expression in a psychic unknown form. The other instance, to which the ego is exposed, originates from the internalization of behavior codes, injunctions, social prohibitions felt as constraint and impediment to the enjoyment of satisfaction, a censorship system that regulates the passage by the instinct from the id to the ego. It is a kind of moral censor able to judge human instinctive acts and desires, based mainly on models of value that the child brings from his relationship with parents, often almost completely unconsciously. Freud named this instance superego.

The concept of psychic function elaborated by Freud seems to be consistent with the latest physiological results on the functional organization of the cerebral cortex. The ego is a mental structure characterized by the function of mediating between the inner world, pulses, impulses, desires from the id, prohibitions from superego and stimuli of external reality by ensuring integration and continuity of the individual. This operating entity identified by Freud finds numerous points of contact today with recent studies coming from the resting state networks of the brain.

As with Freud, many other scientists have tried to explain the organization of thinking apparatus with different theories. The father of American psychology, James (1890), proposed the idea of *stream of consciousness*, underlying how the daily life mental activity flows smoothly with or without the presence of specific stimuli from the external environment. During this state of consciousness, the individual is engaged in recording all the information-bodily sensations (somesthesic and vegetative), experiencing free association of stimuli such as thoughts, memories, past experience, inner dialogue, mental images, emotions, day dreaming, planning future events, and other activities. In this state the mind jumps from one thought to another with fluidity and usually with readiness (Cieri and Esposito, 2018). This state of mind, nowadays called Random Episodic Silent Thinking (REST; Andreasen et al., 1995), emphasizes the free and errant nature of this way of thinking, partly in contrast with the engagement of mind during cognitive tasks.

Among modern neuroimaging techniques, today fMRI and Magnetoencephalography (MEG) allow for the study of the

brain *in vivo*, opening the intersection of anatomy and functions (Cieri and Esposito, 2018). fMRI can be performed during the execution of an experimental paradigm involving specific cognitive tasks or to study spontaneous oscillations of brain activity while the REST of the subject (resting-state fMRI, rs-fMRI; Raichle et al., 2001; Buckner et al., 2008; Cieri and Esposito, 2018; Esposito et al., 2018a). Since it does not require any task, rs-fMRI is characterized as particularly suitable for studies on subjects such as children and elders, because this protocol does not require any particular skill or specific attention focus from the subject, increasing the compliance of the participant and reducing intersubjective variability due to the task performance (Esposito et al., 2018a). Indeed, in recent years a growing number of studies showed that rs-fMRI could be considered as an additional important tool for the investigation of physiological and pathological mental conditions.

Spontaneous brain activity generated in absence of cognitive task has been discussed in the last two decades, representing a pivotal role among psychological and cognitive neuroscientific fields. Many neuroimaging studies considered this brain activity a functioning model of the mind. Cerebral activity recorded during cognitive tasks showed a baseline low frequency fluctuation (0.01–0.1 Hz). In this light some researchers examined these cerebral baseline activities based on the idea that those low levels of brain activity could represent real active states, and that brain activation patterns represent a shift in focus from an internal self-referential state to an external focus (Raichle et al., 2001). This discovery encouraged neuroscientists to begin to consider two different types of neuronal activity: evoked and spontaneous (Fox and Raichle, 2007; Barrett and Simmons, 2015). Brain spontaneous activity has received growing attention (Buckner et al., 2008) in the last decade, supported by several studies showing electric activity, hemodynamic and metabolic parameters, spontaneous fluctuations of membrane potential, spontaneous spikes and neurotransmitter release (O'Donnell and van Rossum, 2014).

During the early 21st century, several studies using PET (Shulman et al., 1997) and task-fMRI (Gusnard and Raichle, 2001) identified specific regions active during cognitive task execution and other brain areas active during different REST conditions. These latter neural regions constituted a network involving both hemispheres: the Medial Prefrontal Cortex (MPFC), the Posterior Cingulate Cortex (PCC), the Inferior Parietal Lobules (IPL), and Hippocampal Regions (HP), forming the neural network called DMN, engaged when mental activity is internally directed, when an individual is left "undisturbed" to think about himself, wondering about his life, his past or future. One hypothesis about the DMN's functioning concerns its involvement in inner mental processes far from all external stimuli, building dynamic mental simulations based on past personal experiences used in recalling memories; it also supports the mental process about the future, and generally when an individual imagines alternative scenarios to the present (Buckner, 2013). This network is also known as a task-negative network because its regions are typically deactivated during execution of attention demanding tasks (Passow et al., 2015).

For many years, modern neuroimaging techniques neglected this important spontaneous activity of the brain, focusing only on changes evoked by external cognitive tasks. During the last two decades, rs-fMRI has become a most utilized tool to study the brain *in vivo*, especially for those patients less cooperative as we mentioned, offering detailed and clear information about the spontaneous brain dynamics in both physiological and pathological conditions (Cieri and Esposito, 2018). Indeed, one of the most important common aims of neuroscience is to identify early biomarkers in order to reach an early diagnosis, providing a timely and specific treatment, even if today we are far from understanding, the neurobiological or neuropsychological markers of all neurological or neuropsychiatric conditions. An important step for neuroscientists and psychoanalysts, useful to reach the mentioned aim to identify early biomarkers, is linked to the deepening of the relationship between mind and brain and mind and body communication. According to Solms (2019), adopting a dual-aspect monist position on the philosophical mind-body problem allows one to find the causal mechanism of consciousness not in the manifest brain but rather in its functional organization, which ultimately underpins both the physiological and the psychological manifestations of experience. Adopting a dual-aspect monist position, using neuroimaging techniques and approaches such as FEP will allow psychoanalysis and neuroscience to investigate this functional organization, studying in deep analysis the mechanisms underlying physiological or pathological human conditions. In this sense and with this common aim, the resting state networks together with the FEP could play a key role in the study of the mind-brain system.

Within this dialogue, the DMN seems to play the same function of mediation attributed by Freud to the ego, and some authors have spoken about Default Self (Beer, 2007; Qin and Northoff, 2011) in order to define the DMN as a kind of biomarker of the Self. Nevertheless, the experience in the perception of the Self is extremely complex, characterized by high variability, and it is not always easy and clear to distinguish the Self from all other phenomena related to cognitive processes. In any case, the role of DMN within the functions of the Self is conspicuous as shown from several psychopathological studies, where the impairment of DMN connectivity associated with an impairment of Self's experience is noticeable.

RESTING STATE NETWORKS IN NEUROPSYCHIATRY

In recent years, there has been a growing interest about abnormal functional connectivity in neurologic and neuropsychiatric disorders, although the results remain debatable. For instance, despite still controversial claims, DMN shows anticorrelated activity with another REST network, the Dorsal Attention Network (DAN), conversely active during externally-directed cognition, such as cognitive tasks that require conscious and focused attention. This anticorrelation could be impaired in some neurological conditions such as Mild Cognitive Impairment (MCI – Esposito et al., 2018a).

Although the focus of this article is not about the use of resting state functional connectivity to assess brain circuits in psychiatric or neurological disorders, it is certainly useful to underline some important issues and connections. Abnormal functional connectivity could be found both in studies on neurodegenerative and neuropsychiatric disorders, including anxiety, major depressive disorder (MDD), bipolar disorder (BD), obsessive compulsive disorder (OCD), schizophrenia (SZ), attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), Eating Behavior Disorder (EBD), Alzheimer's disease (AD), and other neurodegenerative disorders (Buckner, 2013; Andrews-Hanna et al., 2014; Cieri and Esposito, 2018; Esposito et al., 2018a).

In the physiological aging process, the integrity of the DMN is diminished both in function (Andrews-Hanna et al., 2007; Cieri and Esposito, 2018) and structure (Turner and Spreng, 2015), and these changes are associated with MCI, especially in memory functions. Moreover, social cognitive impairments in aging have been associated with reductions in activity within the Dorso Medial Prefrontal Cortex (DMPFC; Moran et al., 2013). These impairments increase in the dimensions of pathological aging as AD and forms of frontotemporal lobar degeneration (FTLD), including semantic dementia (SD) and behavioral variant frontotemporal dementia (bvFTD – Andrews-Hanna et al., 2014).

Schizophrenic patients have shown a dysfunction of an important area of DMN, the Anterior Cingulate Cortex (ACC) associated with difficulty of recognizing actions and functions, correlating with their positive symptoms (Carter et al., 2001). ACC reduces its activity during external cognitive stimulation, highlighting its fundamental role in self-referential mental activity, in close relation with another important region in this process: the anterior insula (AI). Coactivation of these two areas might play a key role in establishing the self-functions (Esposito et al., 2018a,b). In schizophrenic patients, we can observe the typical lack of symbolic ability, lack of activity to make predictions, and mental simulations, often accompanied by the absence of dream activity.

As Andrews-Hanna et al. (2014) pointed out, both the nature and topographical locations of DMN alterations differ across disorders, paralleling varied symptom profiles. While disorders of integrity (e.g., AD) are often associated with hypo-activation or connectivity of a particular DMN component and impairments in specific aspects of self-generated cognition, disorders of content (e.g., depression) and regulation (e.g., ADHD) are typically associated with hyperactivation and hyperconnectivity, paralleled by polarized or excessive forms of self-generated thought (Andrews-Hanna et al., 2014). Moreover, the body image disturbance in EBD may be supported by a modification in connectivity within specific cortical areas like the precuneus (PrC; Seojung et al., 2014), and this result could represent the neural correlates underlying increased self-focus, rumination, and cognitive control in relation to eating disorders and the impairment about the body perception (Esposito et al., 2018a,b).

Within the DMN, specific regions support the self-reported mental processes, monitoring psychological states (Phan et al., 2002) and could be considered regions of convergence receiving internal (bodily and proprioceptive sensations) and external inputs (visual and auditory), for their integration and development. These areas are the cortical midline regions and among these regions the most important are MPFC and ACC, associated with the control of various functions such as selecting or inhibition of some response, monitoring the conflict and identification of errors (Schneider et al., 2008). ACC plays a fundamental role in affective evaluation (Allman et al., 2001), conflict monitoring and detection (Botvinick et al., 2004), response selection (Awh and Jonides, 2001), and attentional control (Posner, 1994).

Andrews-Hanna (2012) and Buckner (2013) hypothesize that one of the major functions of the DMN, perhaps the most important, is to support internal mental simulations used adaptively. This concept is consistent with FEP in which the system is engaged with its simulations, searching of patterns, trying to maintain an internal sensitive balance of the organism, supporting internal mental simulations used in an adaptive way. In other words, the research of patterns claimed by the FEP is consistent with the DMN's most important role, mediating between the external and internal stimuli, building dynamic mental simulations based on past personal experiences used in recalling memories. This is the same function attributed by Freud to the ego (Carhart-Harris and Friston, 2010).

The "investment" of the system in energy terms, in trying to keep lower levels of entropy, decreases the chances of having to face surprises or when optimally attuned to the world, seek out novel situations that will minimize surprise in the future (i.e., expected surprise or uncertainty).

Many neuropsychiatric and neurological diseases are characterized by the impairment or lack of important symbolic function, strictly linked to the Self and to the ability of the system to support internal mental simulations used in adaptive way. Recent findings suggest the existence of a frontoparietal control system consisting of flexible hubs that regulate distributed systems (e.g., visual, limbic, motor) according to current task goals (Cole et al., 2014; Cieri et al., 2017).

DMN seems to directly contribute to all inner mental processes supported by the MPFC and its links to the HP, with its known key role in memory functions. To support this complex interaction, DMN is constituted by two subsystems. The first is the temporal-mesial subsystem, associated with mnemonic processes, activated during retrieval of past memory; this subsystem is predominantly made up of HP and shows high connectivity with another two important brain regions typically active during memory tasks: PCC/PrC (Posterior Cingulate Cortex/Precuneus) and IPL. The second subsystem is connected to the MPFC, specifically dorsal-MPFC activated during mental situations of self-exploration and sensations. The results suggest that self-referential mental activity engages a preferential MPFC subsystem (Szpunar et al., 2007). These functions are closely related to DMN anatomy: two interactive subsystems whose predominant areas are HP and MPFC that converge on the retrosplenial cortex (PCC/PrC).

In the last two decades, rs-fMRI studies have allowed the identification of a set of different networks, not only the DMN, identified in a series of resting-state functional connectivity

studies (Greicius et al., 2003; Fransson, 2005; Fox et al., 2006). In fact, besides the DMN, at least 10 RSN networks have been consistently described in healthy populations (Mantini et al., 2009; van den Heuvel and Hulshoff Pol, 2010; Deco et al., 2011), highlighting that the human brain has a network-based organization at REST. Of these 10, the most studied include the DMN, the Salience Network (SN), the Control Executive Network (CEN) (lateralized in both hemispheres), the primary Sensory Motor Network (SMN), the Extrastriate Visual System (EsV), and the DAN (Deco and Corbetta, 2011).

Important to note in this context is the DAN and its specific behavior related to DMN and the Self. DAN includes Inferior Parietal Sulcus (IPS), Frontal Eye Field (FEF), ACC, and bilateral Middle Temporal Gyrus (MidTempG), and it has received much attention because – conversely to DMN – it is called the task-positive network, being active during cognitive tasks which demand attention and mental control (Corbetta and Shulman, 2002; Fox et al., 2006; Esposito et al., 2018a,b). DMN and DAN show a pattern of anticorrelation in their activity in both task and resting state studies, suggesting that they are intrinsically organized into anticorrelated networks (Fransson, 2005; Esposito et al., 2018a,b). This DAN-DMN anticorrelation during resting state may represent a cerebral mechanism supporting cognitive functions (Gopinath et al., 2015), switching focus between internal, supported by DMN, and external channels and attention demanding events, supported by DAN (Esposito et al., 2018a,b). Interestingly, this negative correlation between DAN and DMN modifies its function during life span. In fact, consistently with function and evolution of the Self, it appears during the first year and it strengthens during the second year of life (Barber et al., 2013).

As mentioned, the concept of Self cannot be seen as a static and steady entity, but rather dynamic in its development and evolution. Developmental psychology claims that a first concept of Self flourishes between the first and second year of life, when the child begins to recognize himself as an object. According to Craig (2011), the most important sign of self-awareness is the ability of the child to recognize himself in the mirror. In parallel, with the growth of the individual Self, the negative correlation between DAN and DMN becomes stronger in adults to support the development of executive functions and working memory from childhood to adulthood (Andrews-Hanna et al., 2007; Cieri and Esposito, 2018).

Following this process, a decreased anticorrelation between these two networks starts to appear weaker during physiological aging (Wu et al., 2011), increasing its weakness in the case of MCI (Esposito et al., 2018a,b), representing a possible biomarker of neuroaging, cognitive decline, and first impairment of self-functions.

DMN AND FREUDIAN SECONDARY PRINCIPLE

Carhart-Harris and Friston (2010) proposed the consistency of the Freudian concept of secondary process with the DMN functions, capable of self-organizing and suppressing free energy, such as the anarchic and unconstrained endogenous activity from the limbic and paralimbic systems. The mind-brain system tries to maintain its state within physiological bounds, trying to minimize the possibility of surprise. This constant attempt to avoid surprise, ensuring that the states remain within physiological bounds, is consistent with the neurophysiological functions of the brain, specifically with the function of DMN.

According to Carhart-Harris and Friston (2010), the construct validity of Freud's hierarchical organization of the mind, with its distinction between id and ego – belonging to the primary and secondary processes, respectively – can be enhanced by remarkable consistency with contemporary models of cognition based on hierarchical Bayesian inference and Helmholtzian free energy. In fact, Freudian metapsychology distinguished two ways of mental functioning, the primary and the secondary processes, corresponding to the pleasure and reality principle, respectively. The primary process is driven by the pleasure principle, which is in turn driven by the id and its instinctual functioning with its instincts and desires, without taking into account the constraints of the external environment with its rules and laws. The secondary process, also called the reality principle, is governed by the ego, which controls the instant gratification mentality of the id. The reality principle is the ability of the mind to assess the reality of external world and to act accordingly with it, as opposed to the pleasure principle.

Freud studied the function of the mind through these different processes, as two fundamentally different styles of cognition, also through a study of non-ordinary states of consciousness (e.g., hallucinations and dreams), in which he recognized a mode of cognition characterized by a primitive style of thinking (Carhart-Harris and Friston, 2010). He speculated that in these primitive non-ordinary states of consciousness, the exchanges of neuronal energy are free, and he designated it as the primary process (Freud, 1940). Moreover, in these non-ordinary states, he identified the loss of certain functions, usually present in "normal" waking cognition, ascribing these functions to a central organization of the ego, which works in order to minimize free energy of the mind, underlying the specific property of this function belonging to the secondary process, defining its aim as one of converting free energy into bound energy states (Carhart-Harris and Friston, 2010).

The Freudian concept of reality principle seems consistent with the functional role of the DMN in its hierarchical and self-organizing role of suppressing free energy originated from subordinate levels, such as the limbic and paralimbic systems. In fact, the Freudian secondary process with its top-down mode of operation, in which it transforms free energy of the lower levels into bound energy trying to keep the system on physiologically acceptable levels, seems to be consistent with the functions of the DMN.

Under this mapping between Freudian and Helmholtzian models, is possible to link the energy associated with the primary process and the free energy of Bayesian formulations; in both accounts, higher cortical areas try to organize the activity from the lower-levels through suppression of their free energy (Carhart-Harris and Friston, 2010).

Another important feature of DMN consistent with FEP is the mentioned anticorrelation, the inverse relationship of its neurophysiological activity with DAN (Corbetta and Shulman, 2002; Fransson, 2005; Fox et al., 2006; Esposito et al., 2018a,b). These intrinsic networks correspond to the high-levels of an inferential hierarchy, which function to suppress the free energy of lower levels (i.e. suppress prediction errors with top-down predictions), associating this optimization process with the Freudian secondary process. Also, the failures of top-down control with non-ordinary states of consciousness, such as early and acute psychosis, the temporal-lobe aura, dreaming, and hallucinogenic drug states (Carhart-Harris and Friston, 2010), might be associated with an impairment of the supraordinate system, as DMN is unable to control in a top-down mode the excess of the free energy from the lower system.

Moreover, as we noticed, the DMN functional connectivity seems to become relatively weak in the elderly (Damoiseaux et al., 2006; Andrews-Hanna et al., 2007), representing a neurological impairment of the mechanism able to support cognitive functions, switching the focus from the inside supported by DMN, to the outside supported by DAN. We can observe the higher control system apparently impaired and unable to bind the free energy, making difficult the executions of cognitive tasks. In cases of ADHD (Castellanos et al., 2008) or impulse control disorders (Church et al., 2009), the hierarchically lower system seems to become too active to be managed by the hierarchically superior system, operating a sort of "mutiny," or "hijacking," leading to an impairment of the system control.

MPFC-PCC connectivity is entirely absent in infants (Fransson, 2005) and the DMN develops through ontogeny, in a way that runs parallel to the emergence of the individual Self with its complex functions.

The spontaneous fluctuations in neuronal activity from cortical nodes of DMN suppress or contain the unconstrained and anarchic endogenous activity of limbic and paralimbic systems (Helmholtz free energy). This neurobiological view rests on the basis of the brain as a hierarchical, inferential, Helmholtzian machine, in which large-scale intrinsic networks such as the DMN are located at higher levels of cerebral hierarchy and work to optimize the representation of the sensorium, minimizing the level of free energy. As Carhart-Harris and Friston (2010) indicate, this optimization, formulated as minimizing free energy, is similar to the treatment of energy in Freudian formulations, and developing these points of contact may help anchor Freudian concepts to more rigorous biological phenomena, helping not only psychoanalysis but the entire neuroscientific field.

As Solms (2014) specifies, when Friston claims about minimizing prediction error and giving up on predictive models that do not correspond to external states, he is making reference to Freud's reality principle, while in a different frame of reference. Freud's descriptions of the secondary process are consistent with the functional anatomy of large-scale intrinsic networks and how this process works to minimize free energy, with its hierarchical organization and continuous and constant attempt trying to keep low levels of surprise. Also, as outlined above, this concordance find is an interesting conceptual hook trough the development of functional connectivity between the nodes of the DMN during ontogeny, as a process that runs parallel to the emergence of the Self's functions.

Freud always explained clinical phenomena in terms of natural forces and energies, not surprisingly he was a student of Helmholtz's medical school and in this regard, it is interesting to note that in 1898 Wilhelm Fliess – an otorhinolaryngologist, passionate scholar of psychoanalysis, and close friend of Freud – sent to him two big volumes of Helmoltz's lessons as a gift in honor of their good friendship and their common attendance and interest in the famous physiologist's lessons and theories.

In the context of the dialogue between psychoanalysis and neuroscience, it might be beneficial for the neuroscience field to try to find contact points with psychoanalysis in order to nourish an inextricable dialogue, started from the birth of psychoanalysis which can certainly improve, providing benefits to the understanding of the mind-brain system in physiological and pathological conditions.

WILFRED RUPRECHT BION: THE THEORY OF "ALPHA FUNCTION"

Freud described the establishment of the principle of reality, underlying how consciousness develops through the perception of the outside world and in addition to the dualism of pleasure-sorrow (the principle of Nirvana – primary narcissism), perception is characterized by manifold sensory qualities (Freud, 1911). Freud's principle of reality is the ability of the mind to assess the outside world, acting accordingly with it in opposed direction to the principle of pleasure (Freud, 1940). Thought is a substitute for motor discharge, even though the latter never stops functioning as a mechanism to release psyche. The establishment of the principle of reality allows the development of a mental function to defer instant gratification, the governing principle of the actions taken by the ego, after its slow development from a "pleasure-ego" into a "reality-ego" (Freud, 1940).

What Freud defined as attention, a mental function that explores outside world, is consistent with Bion's alpha function (α-function), theorized in "Learning From Experience" (Bion, 1962a). In the personality, there are several factors that combined with each other form the personality functions, a term with which Bion intends the mental activity (Bion, 1962a,b). Through α-function, non-mental elements (sensory impressions, β-elements) are processed into mental elements (α-elements), giving them an emotional connotation (good, pleasant, unpleasant, and bad). β-elements are the raw material of mental process, impressions of sensory activation, perceptions of internal and external body state changes that have no meaning and are perceived physically. Everything that is emotionally lived must be at first elaborated by the α-function; this implies that emotional experiences, lived both during sleep and wakefulness, must be elaborated by α-function. When a patient is insufficient in α-function, the β-elements are not thinkable and they can fall under projective identification (Acting-Out). In this case, a patient cannot transform sensory impressions

into α-elements and therefore cannot dream. In order to learn from experience, α-function must operate on the basis of emotional experience by generating α-elements that will be used by thought that works in the dream and in the unconscious (Bion, 1962a, 1973). Dream and α-function are located between conscious and unconscious, differentiating them through a barrier that Bion calls the *contact barrier* that preserves personality from psychotic state. α-function (both awake and during sleep) transforms sensory impressions linked to a specific emotional experience in α-elements that proliferate and condense, forming the contact barrier. The elements can pass freely through the contact barrier between conscious and unconscious states, and the dreams allow us to access directly the contact barrier (Bion, 1962a; Mellor, 2018).

Psychotic patients do not have α-function, resulting in the inability to transform sensorial impressions into α-elements, to dream and to generate conscious and unconscious. In fact, the contact barrier, with its properties necessary to distinguish mental phenomena (conscious and unconscious), is missing in psychotic patients and replaced by the *beta screen* (β-screen) composed of β-elements. Psychotic patients invert α-function, and sensory impressions are no longer used to form α-elements and the contact barrier. α-elements, contact barrier, unconscious thoughts, and dreams are redirected to β-elements and projected to form the β-screen. The inversion of α-function does not recompose β-elements but creates "*bizarre objects*." Indeed, β-elements are sensory impression and do not have traces of personality, while bizarre objects have traces of personality (ego and super-ego). α-function, during the transformation of emotional experience into α-elements, plays a key role in the sense of reality and its inactivity produces disastrous effects on personality such as deep psychotic deterioration.

According to Freud, thoughts are born through the absence, while for Bion thoughts are precedent to thinking, and the latter develops for the necessity to treat thoughts. Bion hypothesized that the mind is a container of thoughts and the α-function develops to contain and process thoughts. It is possible to disengage the mind from thoughts by primitive defense mechanisms, such as expulsion (Freud, 1937), if the personality is prepared to avoid frustration. If, on the other hand, personality is dominated by the impulse to bear and change frustration, it will think the thoughts. If the patient is not able to think his own thoughts, he will have an increase in frustration. Bion underlines that the bear of frustration is a genetically pre-established factor of personality. Models of mental functioning are characterized by the inability to tolerate frustration, suffering, anxiety, and the need to use powerful defense mechanisms: splitting, projection, and projective identification. However, Bion adds that the defense mechanisms concern not only emotions and feelings. Indeed, he proposes a psychotic defense mechanism that splits and free ourselves not only of the intolerable affective content, but of the apparatus that allows its perception, a kind of amputation of specific mind functions. Psychotic defense leads to the impoverishment not only of emotions but also of mental abilities.

All the noted Bion's mind-body unit is in contact with the external reality with the internal sensations supported by α-function. In light of this hypothesis, the mind-body relationship must be seen in continuous dynamism: in harmonic condition, body and mind are integrated with each other, while in disharmonic condition a messy sensoriality that hampers thinking predominates (Ferrari, 1992; Lombardi and Pola, 2010; Lombardi, 2016). In "Transformations" (Bion, 1965), Bion introduces the concept of O. O is the origin as in the geometrical example of the Cartesian axes: experiencing O represents the experience of whole sensations and emotions, which are activated in contact with reality.

As we noted, Francis Joseph Gall (Livianos-Aldana et al., 2007) was the first neuroscientist to study the cerebral cortex, underlying that the brain was made up of several interconnected areas and each of these areas with a specific function. Questioning René Descartes's theory of mind-body dualism, Gall argued that the brain was the seat of intelligence, a theory that was elaborated only after the development of modern psychology. Thanks to Gall's theory, mind was no longer considered separate from body, but as an integral part of the organism in its totality. He noted that the brain was the organ delegated to intellectual, moral and affective faculties, identifying higher psychic functions in the frontal cortex. Empirically, through fMRI studies, cortical midline structures and DMN have often been highlighted to be specific for the Self (Qin and Northoff, 2011). DMN is involved in internally oriented self-related processing that comprises surveillance of internal states (emotional, bodily), resulting in what is called "mind wandering" (Mason et al., 2007). Observing resting state networks in their totality, including the subnetworks (Deco et al., 2011) and their interconnection, we may better understand the mind-body unit. Menon (2011) talks about the "Triple Network," underlying functional interchange between three neural networks: DMN, SN, and CEN. Specifically, DMN with its areas MPFC, PCC, Angular Gyrus, and medial temporal lobe structures plays and important role in monitoring self-referential mental activity; the SN through ACC and Insula, receives and elaborates body sensations and cognitive relevant events engaging frontoparietal systems; CEN, whose key nodes include the dorsolateral prefrontal cortex (DLPFC) and PCC, maintains and elaborates working memory information and decision-making of goal-directed behavior. These networks interact dynamically, mediating cognitive and emotional states (Yu et al., 2018). The SN (Seeley et al., 2007) is involved in bottom-up direction of salience events, involved in detecting, integrating and filtering relevant interoceptive, autonomic and emotional information, and it plays a key role modulating and switching other resting state networks (Menon and Uddin, 2010).

Activation of SN determines an increase of connectivity between DMN and CEN, modulating not only the activation of the networks but also their interconnectivity (Di and Biswal, 2015). SN indeed represents core hubs of the whole brain sending information in other regions and networks. SN through the AI constitutes the hub involved in the registration of internal (body sensations) and external salient

events, sending information to DMN that integrates and elaborates information supporting mental activity connected with the Self (Craig, 2010). AI elaborates affective information, pain and empathy, whereas the dorsal part of ACC (dACC) was most closely associated with conflict resolution and cognitive control. The insula and dACC probably constitute a functional circuit involved in interoceptive and affective processes and form an anatomically tightly coupled network ideally placed to integrate information from several brain regions. The insula distributes sensory information coming from the body, in contact with the external reality, and transmits it to further brain regions that allow its processing. In summary, the insula supports emotional experience resulting from bodily states. In line with Bion's theory, bodily sensations shape emotional experiences, and experiencing O implies the possibility to record the sensations, perceptions and emotions that are activated in contact with reality and therefore experiencing them (Damasio, 1996; Ciocca, 2015). The insula is anatomically situated in a brain area connected with several neural functional circuits supporting cognitive, homeostatic, and affective systems and constitutes a bridge between brain regions involved in monitoring internal states (visceral sensory, somatic sensory processes, autonomic regulation of the gastrointestinal tract and heart; Menon and Uddin, 2010) and that support their processing. The insular cortex registers body sensations and through the interaction with other brain areas, gives rise to emotions that modulate the behavior (Singer et al., 2009; Craig, 2010). Craig and colleagues, in an animal model, identified an ascending pathway from the spinal cord (lamina I neurons in the spinal cord) through the spinothalamic tract, passing through the Nucleus of the Solitary Tract (NTS) and ventromedial nucleus of the thalamus and finally landing at the dorsal insula projecting information to AI and ACC (Critchley and Harrison, 2013). They called this pathway the "*homeostatic afferent pathway*" (Craig, 2009) that carries information about the body. Particularly, information arising from the body reaches the middle and posterior parts of the insula and then is projected in the anterior insula. The awareness of salient events is represented in the anterior insula, whereas more sensory attributes are represented posteriorly (Craig, 2002). The insula represents a core area that receives bodily information, filtering salient stimuli, processing them and then engaging, through ACC, the CEN that supports working memory, higher order cognitive processes and the DMN that supports cognitive functions and Self.

The Freudian description of the mind underlines how bodily experience gives rise to and shapes thought. Pre-reflective representations of visceral states of the Self are linked to activations in the posterior and middle Insula; DMN is engaged when introspection and reflection are needed (Critchley and Harrison, 2013). Interactions between the DMN and insula support the ability to represent one's bodily states to enable conscious reflection on those states (Molnar-Szakacs and Uddin, 2013). Thoughts derive from integrated physiological activation filtered by the insula and he mind develops to process, contain and give them meaning through DMN.

DISCUSSION AND CONCLUSION

The dialogue between neuroscience and psychoanalysis is still complex and often conflicting; a controversy deriving foremost from the complexity of the study object: the mind-brain system, perhaps the most complex and challenging subject for the human being, from a scientific, philosophical, and psychological point of view. A second reason, which probably did not favor the discourse between these two disciplines, derives from the conceptual conflict of conceiving a system able to study itself. Both in the case of neuroscience as in the case of psychoanalysis, the subject and the object of the investigation coincide, and this aspect becomes an evident limitation in the study of any phenomenon. Specifically, these two disciplines use different tools and methods, sharing the same target: the knowledge of the mind-brain system, its development, and its physiological and pathological expressions. The differences in methods and tools used have not discouraged and should not discourage at all this fundamental relationship. Instead, the innovative approach of resting state network investigations has facilitated the communication, opening new horizons. Resting state networks in general and DMN in particular opened a window on neurophysiological mechanisms linked to spontaneous thought processes, not exclusively related to the active execution of cognitive tasks. The greater knowledge of neural networks functioning allows a theoretical deepening on spontaneous and unconscious thought processes and in general on mind-brain functioning and on the mind-body relationship. Although the beginnings of modern neuroscience have been characterized by a cognitive psychology approach, with a tendency to exclude affective, emotional and unconscious processes – in which the unconscious was often defined as implicit or unaware – over time thanks to scientific evidence and clinical practice, it was no longer possible to exclude emotional and unconscious states from neuroscientific studies. This point brought neuroscience back to the approach originally conceived by Freud with the investigation through the resting state networks that confirms and deepens this relationship. Progresses made in the field of neuroimaging allow a deeper and more detailed investigation, therefore a greater understanding of psychoanalytic theory, models and observations, and the mind-brain system in its functions and dysfunctions, finding in concepts such as FEP its natural meeting point, its *bridge* between mind and brain, in which Freud's more speculative free energy theory, based on the clinical method, find a natural connection with the more rigorous methods of neuroscience, a goal to which Freud himself aspired since the birth of psychoanalytic theories.

FEP takes elements from the Bayesian and Helmoltzian approaches, conceiving the human mind as perpetually committed in active inference, analyzing data from the sensorium and from external reality, comparing and analyzing them, trying to keep down the entropy levels and therefore minimize the possibility of surprise (and seeking out opportunities to minimize surprise), thereby avoiding excessive levels of free energy (Friston, 2010). Seth and Friston (2016) recently described active interoceptive inference, providing an interesting and detailed set of concepts within which to conceive the neurofunctional

basis of emotion, embodied selfhood and allostatic control. The neuronal activity encodes expectations about the causes of sensory input, where these expectations aim to minimize prediction error and where the prediction error lies in the difference between (ascending) sensory input and (descending) predictions of that input. This minimization rests upon recurrent neuronal interactions between different levels of the cortical hierarchy. For interoceptive inference, predictions issue from visceromotor areas and project to viscerosensory areas (to provide corollary feedback) as well as to brainstem and subcortical areas (to engage autonomic homoeostatic reflexes). The authors point out how visceromotor predictions are best interpreted as providing homoeostatic set-points that enslave autonomic reflexes and guide allostatic (behavioral and physiological) responses *via* interoceptive prediction errors at different hierarchical levels and timescales (Seth and Friston, 2016).

In the FEP the brain acts having a model of the world, working through active inference generating actively predictions to minimize the variations of free energy (maximizing Bayesian model evidence), providing a principled explanation for perceptual inference in the brain. With this principle, the brain tries to resist its natural tendency to disorder, maintaining a sustained and homoeostatic exchange with the environment.

As we mentioned, the DMN is consistent with ego functions and with its target of containing free energy levels of underlying structures, a function of the secondary process. The result is a top-down hierarchy of DMN which aims to reduce the free energy associated with the Freudian primary process. The cortical regions modulate the activity of subcortical areas, ontogenetically and phylogenetically older, through the lowering and optimization of free energy. Freudian constructs of the primary and secondary processes seem to have neurobiological substrates, consistent with self-organized activity in hierarchical cortical systems, and Freudian descriptions of the ego are consistent with the functions described of the DMN with its reciprocal exchanges with subordinate limbic and paralimbic brain systems.

Even in Bion's theory, the body is in close contact with external reality; internal and external sensations trough α-function shape emotional experiences and finally thoughts. Learning from experience represents the attempt of individuals to experience the emotion of the moment without running away in the knowledge, which would be the result of a defense mechanism aimed at the avoidance of that specific emotional state. The insula, with its connections with several neural functional circuits, supports emotional experience resulting from bodily states.

The anterior insular cortex is a part of the visceromotor area, situated at the top of an interoceptive hierarchy (Seth and Friston, 2016); it receives ascending projections from viscerosensory areas (e.g., posterior and mid-insula) and their descending connections engage a range of subcortical, brainstem, and spinal cord targets involved in visceromotor control, such as the periaqueductal gray and the parabrachial nucleus (Seth and Friston, 2016). The anterior insula constitutes a hub involved in the registration body sensations and filters external salient events, then sending information to the DMN that integrates and elaborates information supporting mental activity connected to the Self. The insula and dACC constitute a functional circuit that integrates information from several brain regions. They form an anatomically tightly coupled network ideally placed to distribute sensory information to further brain regions that allow their processing.

As we noted, Bion (1959) focused his observation on body and sensory organs as instruments of access to the perception of reality, considering thought and emotion inseparable components of the same process, underlying the central role of the body as the start for the thought phenomena. In his theory, digestion of new experiences corresponds to the "data assimilation" or "evidence accumulation" implicit in "belief updating" under the FEP.

Although we do not know if psychoanalysis should help to plan the work of neurobiology, as claimed by Kandel 20 years ago (Kandel, 1999), we believe that a dialogue between these two disciplines should increase in light of new developments, without prejudices in name of curiosity and respect for the history, tools, methodologies, and languages used by the different approaches, in order to reach important advances in the knowledge of the mind-brain system, in which other disciplines as psychiatry, psychology, and neurology could naturally take advantage in order to improve the diagnostic and therapeutic approach to mental suffering.

AUTHOR CONTRIBUTIONS

RE and FC equally contributed to the manuscript.

REFERENCES

Allman, J. M., Hakeem, A., Erwin, J. M., Nimchinsky, E., and Hof, P. (2001). The anterior cingulate cortex: the evolution of an interface between emotion and cognition. *Ann. N. Y. Acad. Sci.* 935, 107–117. doi: 10.1111/(ISSN)1749-6632

Andreasen, N. C., O'Leary, D. S., Cizadlo, T., Arndt, S., Rezai, K., Watkins, G. L., et al. (1995). Remembering the past: two facets of episodic memory explored with positron emission tomography. *Am. J. Psychiatry* 152, 1576–1585. doi: 10.1176/ajp.152.11.1576

Andrews-Hanna, J. R. (2012). The brain's default network and its adaptive role in internal mentation. *Neuroscientist* 18, 251–270. doi: 10.1177/1073858411403316

Andrews-Hanna, J. R., Smallwood, J., and Spreng, R. N. (2014). The default network and self-generated thought: component processes, dynamic control, and clinical relevance. *Ann. N. Y. Acad. Sci.* 1316, 29–52. doi: 10.1111/nyas.12360

Andrews-Hanna, J. R., Snyder, A. Z., Vincent, J. L., Lustig, C., Head, D., Raichle, M. E., et al. (2007). Disruption of large-scale brain systems in advanced aging. *Neuron* 56, 924–935. doi: 10.1016/j.neuron.2007.10.038

Ashby, W. R. (1947). Principles of the self-organising dynamic system. *J. Gen. Psychol.* 37, 125–128. doi: 10.1080/00221309.1947.9918144

Awh, E., and Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends Cogn. Sci.* 5, 119–126. doi: 10.1016/s1364-6613(00)01593-x

Ballard, D. H., Hinton, G. E., and Sejnowski, T. J. (1983). Parallel visual computation. *Nature* 306, 21–26. doi: 10.1038/306021a0

Barber, A. D., Caffo, B. S., Pekar, J. J., and Mostofsky, S. H. (2013). Developmental changes in within- and between-network connectivity between late childhood and adulthood. *Neuropsychologia* 51, 156–167. doi: 10.1016/j.neuropsychologia.2012.11.011

Barrett, L. F., and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* 16, 419–429. doi: 10.1038/nrn3950

Beer, J. S. (2007). The default self: feeling good or being right? *Trends Cogn. Sci.* 11, 187–189. doi: 10.1016/j.tics.2007.02.004

Bion, W. R. (1959). "Attacks on linking" in *Melanie Klein today: Developments in theory and practice. Volume 1: Mainly theory. 1988*. ed. E. Bott Spillius (London: Routledge).

Bion, W. R. (1962a). A theory of thinking. *Int. J. Psychoanal.* 43, 178–186.

Bion, W. R. (1962b). *Learning from experience*. London: Heinemann.

Bion, W. R. (1965). *Transformations*. London: Heinemann.

Bion, W. R. (1973). *Elements of psycho-analysis*. London: Heinemann.

Block, N. (1980). *Readings in philosophy of psychology*. Cambridge, MA: Harvard University Press.

Botvinick, M. M., Cohen, J. D., and Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn. Sci.* 8, 539–546. doi: 10.1016/j.tics.2004.10.003

Buckner, R. L. (2013). The brain's default network: origins and implications for the study of psychosis. *Dialogues Clin. Neurosci.* 15, 351–358.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011

Carhart-Harris, R. L., and Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain* 133, 1265–1283. doi: 10.1093/brain/awq010

Carter, C. S., MacDonald, A. W. 3rd., Ross, L. L., and Stenger, V. A. (2001). Anterior cingulate cortex activity and impaired self-monitoring of performance in patients with schizophrenia: an event-related fMRI study. *Am. J. Psychiatry* 158, 1423–1428. doi: 10.1176/appi.ajp.158.9.1423

Castellanos, F. X., Margulies, D. S., Kelly, C., Uddin, L. Q., Ghaffari, M., Kirsch, A., et al. (2008). Cingulate-precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder. *Biol. Psychiatry* 63, 332–337. doi: 10.1016/j.biopsych.2007.06.025

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Church, J. A., Fair, D. A., Dosenbach, N. U., Cohen, A. L., Miezin, F. M., Petersen, S. E., et al. (2009). Control networks in paediatric Tourette syndrome show immature and anomalous patterns of functional connectivity. *Brain* 132, 225–238. doi: 10.1093/brain/awn223

Cieri, F., and Esposito, R. (2018). Neuroaging through the lens of the resting state networks. *Biomed. Res. Int.* 2018:5080981. doi: 10.1155/2018/5080981

Cieri, F., Esposito, R., Cera, N., Pieramico, V., Tartaro, A., and Di Giannantonio, M. (2017). Late-life depression: modifications of brain resting state activity. *J. Geriatr. Psychiatry Neurol.* 30, 140–150. doi: 10.1177/0891988717700509

Ciocca, A. (2015). *Storia della psicoanalisi*. Bologna: Il Mulino.

Cole, M. W., Repovs, G., and Anticevic, A. (2014). The frontoparietal control system: a central role in mental health. *Neuroscientist* 20, 652–664. doi: 10.1177/1073858414525995

Connolly, J. P. (2016). *Principles of organization of psychic energy within psychoanalysis: a systems theory perspective*. Unpublished doctoral thesis. Pretoria: University of South Africa.

Connolly, P. (2018). Expected free energy formalizes conflict underlying defense in Freudian psychoanalysis. *Front. Psychol.* 9:1264. doi: 10.3389/fpsyg.2018.01264

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nat. Rev. Neurosci.* 3, 655–666. doi: 10.1038/nrn894

Craig, A. D. (2009). How do you feel – now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10, 59–70. doi: 10.1038/nrn2555

Craig, A. D. (2010). The sentient self. *Brain Struct. Funct.* 214, 563–577. doi: 10.1007/s00429-010-0248-y

Craig, A. D. (2011). Significance of the insula for the evolution of human awareness of feelings from the body. *Ann. N. Y. Acad. Sci.* 1225, 72–82. doi: 10.1111/j.1749-6632.2011.05990.x

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

Critchley, H. D., and Harrison, N. A. (2013). Visceral influences on brain and behavior. *Neuron* 77, 624–638. doi: 10.1016/j.neuron.2013.02.008

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 351, 1413–1420.

Damoiseaux, J. S., Rombouts, S. A., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., et al. (2006). Consistent resting-state networks across healthy subjects. *Proc. Natl. Acad. Sci. USA* 103, 13848–13853. doi: 10.1073/pnas.0601417103

Dayan, P., Hinton, G. E., and Neal, R. M. (1995). The Helmholtz machine. *Neural Comput.* 7, 889–904. doi: 10.1162/neco.1995.7.5.889

Deco, G., and Corbetta, M. (2011). The dynamical balance of the brain at rest. *Neuroscientist* 17, 107–123. doi: 10.1177/1073858409354384

Deco, G., Jirsa, V. K., and McIntosh, A. R. (2011). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.* 12, 43–56. doi: 10.1038/nrn2961

Di, X., and Biswal, B. B. (2015). Characterizations of resting-state modulatory interactions in the human brain. *J. Neurophysiol.* 114, 2785–2796. doi: 10.1152/jn.00893.2014

Esposito, R., Cieri, F., Chiacchiaretta, P., Lauriola, M., Di Giannantonio, M., Tartaro, A., et al. (2018a). Modifications in resting state functional anticorrelation between default mode network and dorsal attention network: comparison among young adults, healthy elders and mild cognitive impairment patients. *Brain Imaging Behav.* 12, 127–141. doi: 10.1007/s11682-017-9686-y

Esposito, R., Cieri, F., di Giannantonio, M., and Tartaro, A. (2018b). The role of body image and self-perception in anorexia nervosa: the neuroimaging perspective. *J. Neuropsychol.* 12, 41–52. doi: 10.1111/jnp.12106

Ferrari, A. B. (1992). *L'eclissi del corpo*. Borla: Roma.

Fox, M. D., and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* 8, 700–711. doi: 10.1038/nrn2201

Fox, M. D., Snyder, A. Z., Zacks, J. M., and Raichle, M. E. (2006). Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nat. Neurosci.* 9, 23–25. doi: 10.1038/nn1616

Fransson, P. (2005). Spontaneous low-frequency BOLD signal fluctuations: an fMRI investigation of the resting-state default mode of brain function hypothesis. *Hum. Brain Mapp.* 26, 15–29. doi: 10.1002/hbm.20113

Freud, S. (1940). *An outline of psychoanalysis. Std Edn. Vol. 23*. London: Vintage.

Freud, S. (1911). *Formulations on the two principles of mental functioning. Std Edn. Vol. 12.* 218–286.

Freud, S. (1895/1963). "On the grounds for detaching a particular syndrome from neurasthenia under the description 'anxiety neurosis'" in *The standard edition of the complete psychological works of Sigmund Freud. Vol. 3,* ed. J. Strachey trans. (London: The Hogarth Press).

Freud, S. (1895/1966). "Project for a scientific psychology" in *The standard edition of the complete psychological works of Sigmund Freud. Vol. 1,* ed. J. Strachey trans. (London: The Hogarth Press).

Freud, S. (1915). *Instincts and their vicissitudes. The standard edition of the complete psychological works of Sigmund Freud, volume XIV (1914–1916): On the history of the psycho-analytic movement, papers on metapsychology and other works.* 109–140.

Freud, S. (1923). "The ego and the Id" in *The standard edition of the complete psychological works of Sigmund Freud: The ego and the Id and other works*. Vol. 19, eds. J. Strachey, A. Freud, A. Strachey and A. Tyson, (London: Vintage), 1–66.

Freud, A. (1937). *The ego and the mechanisms of defence*. London: Hogarth Press and Institute of Psycho-Analysis.

Friston, K. J. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 360, 815–836.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle of the brain. *J. Physiol.* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Friston, K., Levin, M., Sengupta, B., and Pezzulo, G. (2015a). Knowing one's place: a free-energy approach to pattern regulation. *J. R. Soc. Interface* 12, 1–12. doi: 10.1098/rsif.2014.1383

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015b). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi: 10.1080/17588928.2015.1020053

Gopinath, K., Krishnamurthy, V., Cabanban, R., and Crosson, B. A. (2015). Hubs of anticorrelation in high-resolution resting-state functional

connectivity network architecture. *Brain Connect.* 5, 267–275. doi: 10.1089/brain.2014.0323

Gregory, R. L. (1980). Perceptions as hypotheses. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 290, 181–197.

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. USA* 100, 253–258. doi: 10.1073/pnas.0135058100

Gusnard, D. A., and Raichle, M. E. (2001). Searching for a baseline: functional imaging and the resting human brain. *Nat. Rev. Neurosci.* 2, 685–694. doi: 10.1038/35094500

Helmholtz, H. (1866/1962). "Concerning the perceptions in general" in *Treatise on physiological optics. 3rd Edn. Vol. III*. ed. J. Southall trans. (New York: Dover).

Hopkins, J. (2012). "Psychoanalysis, representation and neuroscience: the Freudian unconscious and the Bayesian brain" in *From the couch to the lab: Psychoanalysis, neuroscience and cognitive psychology in dialogue*. eds. A. Fotopoulu, D. Pfaff and M. Conway (Oxford: Oxford University Press), 230–265.

Hopkins, J. (2015). "The significance of consilience: psychoanalysis, attachment, neuroscience, and evolution" in *Psychoanalysis and philosophy of mind: Unconscious mentality in the 21st century*. eds. S. Boag, L. Brakel and V. Talvitie (London: Karnac), 47–136.

Hopkins, J. (2016). Free energy and virtual reality in neuroscience and psychoanalysis: a complexity theory of dreaming and mental disorder. *Front. Psychol.* 7:922. doi: 10.3389/fpsyg.2016.00922

James, W. (1890). "The principles of psychology" (London: MacMillan).

Kandel, E. (1999). Biology and the future of psychoanalysis: a new intellectual framework for psychiatry revisited. *Am. J. Psychiatry* 156, 505–524.

Kauffman, S. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford: Oxford University Press.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15:20170792. doi: 10.1098/rsif.2017.0792

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Levin, K. (1935). "Dynarnic theory of personality" in *Psychology* ed. J. F. Dashiell (New York and London: McGraw-Hill Book Company, Inc.).

Livianos-Aldana, L., Rojo-Moreno, L., and Sierra-Sanmiguel, P. F. J. (2007). Gall and the phrenological movement. *Am. J. Psychiatry* 164:414. doi: 10.1176/ajp.2007.164.3.414

Lombardi, R. (2016). *Metà prigioniero, metà alato. La dissociazione corpo-mente in psicoanalisi*. Bollati Boringhieri: Torino.

Lombardi, R., and Pola, M. (2010). The body, adolescence, and psychosis. *Int. J. Psychoanal.* 91, 1419–1444. doi: 10.1111/j.1745-8315.2010.00356.x

Lurija, A. R. (1976). *Working brain: An introduction to neuropsychology*. ISBN 046509208X (ISBN13: 9780465092086).

Mantini, D., Corbetta, M., Perrucci, M. G., Romani, G. L., and Del Gratta, C. (2009). Large-scale brain networks account for sustained and transient activity during target detection. *NeuroImage* 44, 265–274. doi: 10.1016/j.neuroimage.2008.08.019

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., and Macrae, C. N. (2007). Wandering minds: the default network and stimulus-independent thought. *Science* 315, 393–395. doi: 10.1126/science.1131295

McEwen, B. S. (2004). "Protective and damaging effects of the mediators of stress and adaptation: allostasis and allostatic load" in *Allostasis, homeostasis, and the costs of physiological adaptation*. ed. J. Schulkin (Cambridge, MA: Cambridge University Press), 65–98.

McEwen, B. S. (2007). Physiology and neurobiology of stress and adaptation: central role of the brain. *Physiol. Rev.* 87, 873–904. doi: 10.1152/physrev.00041.2006

Mellor, M. J. (2018). Making worlds in a waking dream: where Bion intersects Friston on the shaping and breaking of psychic reality. *Front. Psychol.* 9:1674. doi: 10.3389/fpsyg.2018.01674

Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn. Sci.* 15, 483–506. doi: 10.1016/j.tics.2011.08.003

Menon, V., and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* 214, 655–667. doi: 10.1007/s00429-010-0262-0

Miller, J. G. (1978). *Living systems*. New York: McGraw-Hill.

Molnar-Szakacs, I., and Uddin, L. Q. (2013). Self-processing and the default mode network: interactions with the mirror neuron system. *Front. Hum. Neurosci.* 7:571. doi: 10.3389/fnhum.2013.00571

Moran, L. V., Tagamets, M. A., Sampath, H., O'Donnell, A., Stein, E. A., Kochunov, P., et al. (2013). Disruption of anterior insula modulation of large-scale brain networks in schizophrenia. *Biol. Psychiatry* 74, 467–474. doi: 10.1016/j.biopsych.2013.02.029

O'Donnell, C., and van Rossum, M. C. (2014). Systematic analysis of the contributions of stochastic voltage gated channels to neuronal noise. *Front. Comput. Neurosci.* 8:105. doi: 10.3389/ fncom. 2014. 00105

Ortega, P. A., and Braun, D. A. (2010). A minimum relative entropy principle for learning and acting. *J. Artif. Intell. Res.* 38, 475–511.

Passow, S., Specht, K., Adamsen, T. C., Biermann, M., Brekke, N., Craven, A. R., et al. (2015). Default-mode network functional connectivity is closely related to metabolic activity. *Hum. Brain Mapp.* 36, 2027–2038. doi: 10.1002/hbm.22753

Phan, K. L., Wager, T., Taylor, S. F., and Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage* 16, 331–348. doi: 10.1006/nimg.2002.1087

Posner, M. I. (1994). Attention: the mechanisms of consciousness. 1994. *Proc. Natl. Acad. Sci. USA* 91, 7398–7403.

Qin, P., and Northoff, G. (2011). How is our self related to midline regions and the default-mode network? *NeuroImage* 57, 1221–1233. doi: 10.1016/j.neuroimage.2011.05.028

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. USA* 98, 676–682. doi: 10.1073/pnas.98.2.676

Ramstead, M. J. D., Badcock, P. J., and Friston, K. J. (2018). Answering Schrödinger's question: a free-energy formulation. *Phys. Life Rev.* 24, 1–16. doi: 10.1016/j.plrev.2017.09.001

Rapaport, D. (1960). The structure of psychoanalytic theory. *Psychol. Issues* 2, 1–158.

Schneider, F., Bermpohl, F., Heinzel, A., Rotte, M., Walter, M., Tempelmann, C., et al. (2008). The resting brain and our self: self-relatedness modulates resting state neural activity in cortical midline structures. *Neuroscience* 157, 120–131. doi: 10.1016/j.neuroscience.2008.08.014

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., et al. (2007). Dissociable Intrinsic Connectivity Networks for Salience Processing and Executive Control. *J. Neurosci.* 27, 2349–2356. doi: 10.1523/JNEUROSCI.5587-06.2007

Seojung, L., Kyung, B. K., Jeonghun, K., Jung-Hyun, L., Kee, N., and Young-Chul, J. (2014). Resting-state synchrony between anterior cingulate cortex and precuneus relates to body shape concern in anorexia nervosa and bulimia nervosa. *Psychiatry Res.* 221, 43–48. doi: 10.1016/j.pscychresns.2013.11.004

Seth, A. K., and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Phil. Trans. R. Soc. B* 371:20160007. doi: 10.1098/rstb.2016.0007

Shulman, G. L., Fiez, J. A., Corbetta, M., Buckner, R. L., Miezin, F. M., Raichle, M. E., et al. (1997). Common blood flow changes across visual tasks: II. Decreases in cerebral cortex. *J. Cogn. Neurosci.* 9, 648–663. doi: 10.1162/jocn.1997.9.5.648

Singer, T., Critchley, H. D., and Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends Cogn. Sci.* 13, 334–340. doi: 10.1016/j.tics.2009.05.001

Solms, M. (2014). A neuropsychoanalytical approach to the hard problem of consciousness. *J. Integr. Neurosci.* 13, 173–185. doi: 10.1142/S0219635214400032

Solms, M. (2019). The hard problem of consciousness and the free energy principle. *Front. Psychol.* 9:2714. doi: 10.3389/fpsyg.2018.02714

Sterling, P. (2004). "Principles of allostasis: optimal design, predictive regulation, pathophysiology, and rational therapeutics" in *Allostasis, homeostasis, and the costs of physiological adaptation*. ed. J. Schulkin (Cambridge, MA: Cambridge University Press).

Sterling, P., and Eyer, J. (1988). "Allostasis: a new paradigm to explain arousal pathology" in *Handbook of life stress, cognition, and health*. eds. S. Fisher and J. Reason (Chichester, UK: John Wiley and Sons), 629–649.

Szpunar, K. K., Watson, J. M., and McDermott, K. B. (2007). Neural substrates of envisioning the future. *Proc. Natl. Acad. Sci. USA* 104, 642–647. doi: 10.1073/pnas.0610082104

Tretter, F., and Löffler-Stastka, H. (2018). Steps toward an integrative clinical systems psychology. *Front. Psychol.* 9:1616. doi: 10.3389/fpsyg.2018.01616

Turner, G. R., and Spreng, R. N. (2015). Prefrontal engagement and reduced default network suppression co-occur and are dynamically coupled in older adults: the default-executive coupling hypothesis of aging. *J. Cogn. Neurosci.* 27, 2462–2476. doi: 10.1162/jocn_a_00869

van den Heuvel, M. P., and Hulshoff Pol, H. E. (2010). Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* 20, 519–534. doi: 10.1016/j.euroneuro.2010.03.008

von Bertalanffy, L. (1967). *General system theory: Foundations, development, applications*. New York: George Braziller.

von Helmholtz, H. (1962). "Concerning the perceptions in general" in *Treatise on physiological optics, 3rd Edn. Vol. III* (translated by J. P. C. Southall 1925 Opt. Soc. Am. Section 26, reprinted New York: Dover, 1962).

Wu, J. T., Wu, H. Z., Yan, C. G., Chen, W. X., Zhang, H. Y., He, Y., et al. (2011). Aging-related changes in the default mode network and its anti-correlated networks: a resting-state fMRI study. *Neurosci. Lett.* 504, 62–67. doi: 10.1016/j.neulet.2011.08.059

Yu, Y., Yang, J., Ejima, Y., Fukuyama, H., and Wu, J. (2018). Asymmetric functional connectivity of the contra- and ipsilateral secondary somatosensory cortex during tactile object recognition. *Front. Hum. Neurosci.* 11:662. doi: 10.3389/fnhum.2017.00662

Zepf, S. (2010). Libido and psychic energy – Freud's concepts reconsidered. *Int. Forum Psychoanal.* 19, 3–14. doi: 10.1080/08037060802450753

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 Psychology

HYPOTHESIS AND THEORY
published: 26 September 2019
doi: 10.3389/fpsyg.2019.02173

Working With the Predictable Life of Patients: The Importance of "Mentalizing Interoception" to Meaningful Change in Psychotherapy

Patrice Duquette[1]* *and Vivien Ainley*[2]

[1] Private Practice, Birmingham, MI, United States, [2] Lab of Action and Body, Royal Holloway, University of London, Egham, United Kingdom

To understand our patients and optimize their treatment, psychotherapists of all theoretical orientations may benefit from considering current scientific evidence alongside psychodynamic constructs. There is recent neuroscientific evidence that subjective awareness, feelings and emotions depend upon "interoception," defined as the neural signaling to the brain from all tissues of the body. Interoception is the obvious basis of homeostasis (in the brainstem) but some interoceptive signals rise above this level and contribute to inferential processes that substantiate intrapersonal and interpersonal experience. The focus of this paper is on the essential role that their "interoception" plays in our patients' emotional experience and subjective awareness, and how the process referred to as "mentalizing interoception" may be harnessed in therapy. This can best be understood in terms of "predictive processing," which describes how subjective states, and particularly emotion, are inferred from sensory inputs – both interoceptive and exteroceptive. Predictive processing assumes that the brain infers (probabilistically) the likely cause of sensation experienced through the sense organs, by testing this sensory data against its innate and learned "priors." This implies that any effort at changing heavily over-learned prior beliefs will require action upon the system that has generated that set of prior beliefs. This involves, quite literally, acting upon the world to alter inferential processes, or in the case of interoceptive priors, acting on the patient's body to alter habitual autonomic nervous system (ANS) reflexes. Focused attention to bodily sensations/reactions, in the safety of the therapeutic relationship, provides a route to "mentalizing interoception," by means of the bodily cues that may be the only conscious element of deeply hidden priors and thus the clearest way to access them. This can: update patients' characteristic, dysfunctional responses to emotion and feelings; increase emotional insight; decrease cognitive distortions; and engender a more acute awareness of the present moment. These important ideas are outlined below from the perspective of psychodynamic psychotherapeutic practice, in order to discuss how relevant information from neuroscientific theory and current research can best be applied in clinical treatment. A clinical case will be presented to illustrate how this argument or treatment relates directly to clinical practice.

Keywords: interoception, mentalization of interoception, emotion, predictive processing, free energy principle, psychotherapy, psychotherapeutic change

OPEN ACCESS

Edited by:
Christoph Mathys, International School for Advanced Studies (SISSA), Italy

Reviewed by:
Karl Friston, University College London, United Kingdom
Michael Moutoussis, University College London, United Kingdom

***Correspondence:**
Patrice Duquette
patrice@pmdmd.com

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *12 March 2019*
Accepted: *09 September 2019*
Published: *26 September 2019*

Citation:
Duquette P and Ainley V (2019) Working With the Predictable Life of Patients: The Importance of "Mentalizing Interoception" to Meaningful Change in Psychotherapy. Front. Psychol. 10:2173. doi: 10.3389/fpsyg.2019.02173

INTRODUCTION

Patients undergoing psychotherapeutic work, in any theoretical orientation, often struggle to identify and verbalize their emotional experiences and to explain their subjective views of the world, in the face of the confusing messages arising from within their physical bodies, which are accompanied by their persistent, strongly-held beliefs about themselves and the world. A common experience brings individuals to our office:

> "*It seems that people spend most of their time with the delusion that they have an accurate representation of the world. Actually, evidence suggests that we are all rather poor at letting our sensory experience update our beliefs, and that we are susceptible to prior beliefs and social constraints that greatly limit our ability to deal with evidence rationally. For most of us, this may be manifest. . .. as vulnerability to biases as we try to model the world*" (Fletcher and Frith, 2009, p. 52).

At any given moment, physical sensations can lead the way into a distorted view of ourselves and of reality. Emotions, especially of fear, can dominate subjective experience; biasing our assessment of what is truly emotive in our bodies and what actual meaning this may have in the current moment. Prior beliefs or expectations stimulate reactive processes, quickly defining subjective experience, allowing little room for any testing of these potentially distorted beliefs against reality. For example: if the patient's heart races, they tend to believe they must be scared and that they face real danger; if they have tears in their eyes they claim this is the result of some hurt and another person is responsible. Whether they are trying to: describe emotional states; parse out elements of experience; or ascertain the reality of events and interactions versus those they imagine, patients often struggle to constrain the habitual influence that their body has on these processes.

Such rapid knee jerk reactions to stimuli are learned in early infancy, where all sensation feels forever and is mostly a surprise. We are "*born too early*" (Bar-Levav, 1988), such that the processes of bodily experiences that shape emotional processes are initiated before the world can be comprehended or tested against reality. Crucially, as the body changes and adapts, the brain is simultaneously establishing expectations about relationships and the environment, within and without. These early embodied patterns persist throughout life, strongly influencing how individuals understand, behave and experience the world, intra- and inter-subjectively. What our patients know of themselves and the world is "in their bones" – acquired in childhood from experience comprised of motoric, humoral, neural, sensory and autonomic responses to salient stimuli.

Neuroscience has recently increased our knowledge of the vital processes that send neural information from the body to the brain – regulating life processes at basic levels, while also modulating emotional experience and subjective awareness at the most complex mental levels. This process is termed "interoception," which refers to the constant flow of signals passing between the body and the brain that are the "*foundation for the sequential integration of your homeostatic condition with your sensory environment, with your motivational condition, and with your social condition*" (Craig, 2008, p. 281).[1] At an elemental level, interoception instantiates physiologic homeostatic regulation of the body and is part of the neural infrastructure that determines emotional experience and subjective awareness, ultimately influencing cognition and behavior. When theorists claim that the processing of interoceptive information from the body underpins the processes of emotion generation, feelings and affect, they are talking literally about gut feelings.

It is now recognized that interoceptive signals, combined with information from other exteroceptive sensory modalities (like vision and touch) are integrated with top-down learned expectations in the brain, thus contributing not only to homeostasis but crucially to emotion – thus ultimately influencing cognition and behavior (Critchley and Harrison, 2013). The rubric of "predictive processing" is a valuable model within which to consider current research and possible therapy. The basic premise in predictive processing is that humans do not have direct access to the truth about our internal and external environment but that our brains must make inferences about these, on the basis of the sensory evidence that we have (Friston, 2010, 2013; Pezzulo, 2014; Clark, 2016). This approach stresses that the brain's task is to minimize the difference between the actual incoming sensory data and what the brain expects or infers (i.e., "predicts" from experience) as the most likely cause of whatever the sensory organs are currently registering (Edwards et al., 2012; Barrett et al., 2016; Seth and Friston, 2016; Critchley and Garfinkel, 2017).

At the leading edge of research, the over-arching principle of "free energy" accounts for how inferential processes support humans' inherent drive toward homeostasis and self-organization, by minimizing uncertainty, which is defined as the difference between the actual states of the body and the states the brain infers are optimal for its Darwinian success (Friston, 2010). The insights that these new ideas offer about how individuals experience themselves, other people and the world in general adds meaningful dimensions to therapeutic practice that were unavailable even two decades ago.

"Interoceptive inference" is the specific aspect of predictive processing which refers to how we interpret internal sensations (Gu and FitzGerald, 2014; Ondobaka et al., 2017; Owens et al., 2018; Allen et al., 2019). Much of interoceptive signaling is unconscious, or at the very borders of awareness (Adam, 2010), involving the pre-reflective, sub-personal assimilation of interoceptive bodily cues. However, implicit contextualization of autonomic reflexes, and reactions to emotionally salient cues, occurs in the body all the time. These processes are generally not available to awareness but they nevertheless have powerful impact on emotions and feelings states and also on behavior – potentially leading to persistent difficulty in our patient's lives.

[1] Researchers in neuroscience have studied interoception from a variety of vantage points which has resulted in a range of definitions and perspectives. Discussion of different facets of interoception are beyond the scope of this paper but see Ceunen et al. (2016), Khalsa and Lapidus (2016), and Duquette (2017) for overviews of how the study of interoception has changed over the last century.

In this paper, we propose that the "mentalization" of interoceptive sensations is elemental in making these pre-reflective processes available for self-reflection. The verbalization or expression of what the patient finds on self-reflection is the important starting point between patient and therapist – providing language for feeling states and bringing emotion to the level of subjective experience – which is an important goal for any psychological school of thought.

We refer to this process as "mentalizing interoception." While the term "mentalizing" is commonly used to denote inferring or understanding the mental states of others it also refers crucially to mental states of the self (e.g., Fonagy and Target, 2006). We use the term "mentalizing" here with the ultimate goal that the patient will have an "*intentional mental state*" (Fonagy and Target, 2006; Allen et al., 2008; Bateman and Fonagy, 2012), our usage differs in that we are assuming that the term "mentalization" specifically includes inferring the imagined causes and implications of sensation that impact on the individual (Besharati et al., 2015; Fotopoulou, 2015; Fotopoulou and Tsakiris, 2017).

Following Fotopoulou and Tsakiris (2017), we assume that the backdrop to mentalizing interoception (which they also call "*embodied mentalization*") is the "*on-going, dynamic process of maintaining and updating generative models of the likely cause of sensory data from inside the body itself and the external world*" (Fotopoulou and Tsakiris, 2017). However, while Fotopoulou and Tsakiris (2017) are principally concerned with how the infant's development of the experience of the self requires mentalization of interoception, we focus here on the mentalization of interoception as the ongoing process of intentional, self-reflective evaluation of interoceptive sensation that can occur in the immediate present for the adult patient.

The crux of our argument is the proposal that attention to interoceptive sensation can be harnessed in therapy to support change, given the essential role that interoception plays in our patients' emotional experience and their subjective awareness. Specifically, our purpose is to show how the patients' mentalization of interoception can lead to the generation of newly imagined possibilities regarding current interoceptive sensations, thus bringing ongoing interoceptive inferences into awareness at a self-reflective level. Within relational interactions with the therapist, the patient can then test their sub-optimal but habitual prior beliefs, about themselves and the world (and their consequent emotions and behaviors) and create alternative, more flexible opportunities for experience and action.

Within any clinical approach, a great deal can be gained if psychotherapists comprehend: the full significance of interoception and physiological regulatory processes for subjective experience; the power of inferential processes in how we all make meaning of our sensory world; and our reliance on habitual reactions as we try to limit the uncertainty that is inherent in human experience. Understanding processes that are constantly active but often only evidenced in bodily signatures can inform and anchor the therapeutic interaction, as the patient engages in the task of generating new hypotheses (and corresponding language) and thereby creating alternative perspectives to loosen the bounds of long-held, over-determined ways of seeing, relating, and behaving in their world. We will bring to the fore current theory and research that is most relevant for practicing clinicians and will consider how these ideas can add to their practice.

Firstly, we briefly describe the neurobiology of interoception and its place in homeostatic and allostatic regulation and thus in physiologic stability, together with an outline of the embodied (interoceptive) nature of emotion and subjective experience. All-important to our argument is the manner in which an individual's model of the world is shaped by the interaction of their interoceptive signals with higher order inferences, in the form of early (unconscious) "prior beliefs" that may be partly innate but are also learned. Crucially, these inferential processes depend on the minimization of uncertainty through prediction. Such prior beliefs have great potential for distortion and we suggest means by which therapists can identify the state of the patient's interoceptive inferential processes and hence gain insight into their health and psychopathology. Possible interventions are suggested whereby clinicians may encourage meaningful introspection and emotional openness in patients, while relationally supporting their efforts to alter long-held perspectives about their embodied experience and its effect on their view of themselves and the world.

It is hoped that applying the lessons from predictive processing and free energy to the therapeutic process will encourage conversations across theoretical lines, while increasing collaboration between researchers in neuroscience and psychology.

INTEROCEPTION AND PHYSIOLOGIC REGULATION

We learn in childhood that we have five senses with which to experience the world (and ourselves within it) but this classic account neglects that we also perceive the world through sensations generated from within our own body, where every cell contributes to our experience, i.e., through interoception. Memories and learned associations contribute to this process (Craig, 2002, 2008; Critchley and Harrison, 2013; Ceunen et al., 2016; Khalsa and Lapidus, 2016). The interoceptive pathway originates in cells in all types of tissues of the body – including muscles, joints, teeth, skin and all the viscera (Craig, 2002) and these interoceptive signals flow to the brain through designated neural fibers (Critchley and Harrison, 2013) (see **Box 1**).

Interoception is functionally fundamental to homeostasis, which is largely determined unconsciously, countering the inherent instability of the organism and maintaining internal physiologic order amidst the stressor of the ever-changing external environment (Cannon, 1932). Importantly, interoceptive signals produce sensations which are experienced as pleasant or unpleasant, creating motivation within the individual (consciously or not) to move toward or away from the sensation (Craig, 2008, 2010; Duquette, 2017). The effect of this is that the flow of interoceptive signals motivates the behavior that

BOX 1 | Neuroanatomy of interoception.

The neural pathway of interoception originates in the various tissues of the body in small diameter fibers (A and C-delta type) which transmit neural signals regarding pain, temperature, blood osmolality, and metabolic needs. These include nociceptors, thermoreceptors, osmoreceptors, and metaboreceptors. The afferent fibers collecting neuronal signals from receptors within the body transmit neural signals to lamina I – a layer of tissue that extends up through the spinal cord to the brain ("Afferent" = from the body to the brain, "efferent" = from the brain to the body). Within the brainstem, fibers carrying interoceptive information interact extensively with both branches of the autonomic nervous system (ANS), allowing a nearly instantaneous response to interoceptive neural information and thus heightening homeostatic autonomic control (Craig, 2008). Spreading into the brain, small diameter fibers project to multiple neuroanatomic areas including: nuclei within the periaqueductal gray (PAG); the parabrachial nucleus (PBN); the nucleus of the solitary tract (NTS); the thalamus (notably the Ventromedial Nucleus); and insular cortex (IC).

The IC is a cortical area that is deeply folded and set within the large sulcus, or groove, of between the frontal and temporal lobes (see **Figure 1**). Neuronal signals progress through the different sections which have different cellular architecture and functional purposes; these are the posterior, middle and anterior insular cortices. For greater detail on the functional purposes of the different insular cortical sections see **Box 2**.

The insula is an important brain hub with wide interconnections. As well as integrating interoceptive signals, the insula receives direct input from the exteroceptive sensory cortex (for sensation from external organs, e.g., hearing, vision etc.) (Northoff, 2016). There are bidirectional connections from the insula to several areas, the prefrontal cortex, parietal and temporal cortex, basal ganglia, with the connections to the anterior cingulate cortex (ACC) heavily studied and elucidated.

The ACC plays an important complementary role to the anterior insula cortex (AIC). Most researchers agree that the AIC and ACC serve as interdependent arms of a coordinated system, which has been described as "limbic sensory" (AIC) and "limbic motor" (ACC) cortices (Craig, 2009b). While the anterior insula is assumed to underpin all feelings and awareness, the anterior cingulate is related more to motivation and behavior (Medford and Critchley, 2010; Craig, 2011). The coordinated function of the AIC and the ACC creates integrated awareness of our cognitive, affective and physical state, which then serves as the basis for the selection of, and preparation for, our responses to internal and external events (Medford and Critchley, 2010).

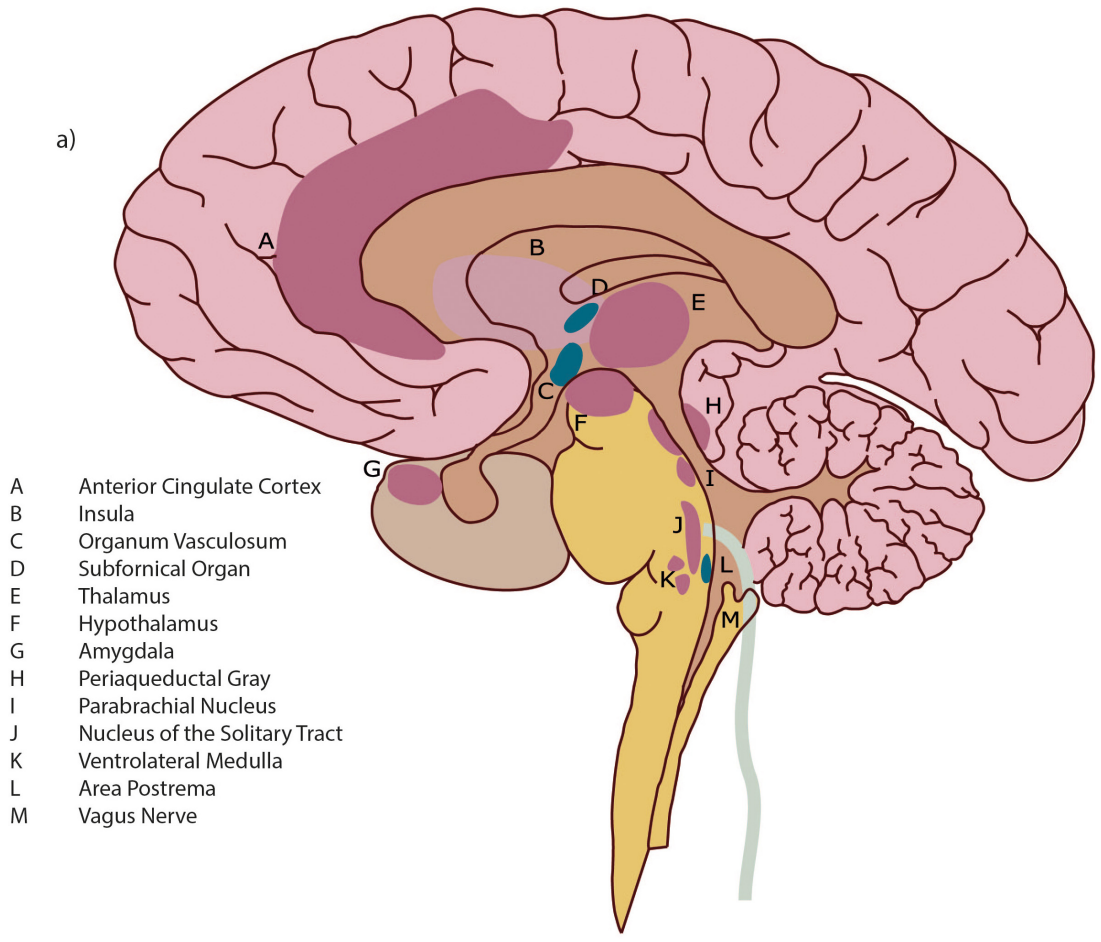

FIGURE 1 | Detail of brain architecture. Reprinted with permission from Quadt et al. (2018) Annals of the New York Academy of Sciences.

is necessary to maintain homeostatic equilibrium – hence the essential role of interoception in motivated action (Craig, 2002, 2009b; Strigo and Craig, 2016; Critchley and Garfinkel, 2017). While homeostasis has previously been characterized as primarily reactive control of physiologic states, the concept of "allostatic" regulation is more relevant over any extended time period, as

this invokes prospective control, to avoid problematic deviations from homeostatic set-points (Sterling, 2012, 2014; Stephan et al., 2016; Petzschner et al., 2017).

Allostasis, defined as "*stability through change*," (Sterling, 2014, p. 1192) utilizes interoceptive information to implement control of bodily states in order to anticipate energy demands in advance of perturbations which would otherwise be dangerous (Sterling, 2012, 2014; Barrett et al., 2016; Stephan et al., 2016). Allostatic mechanisms are implemented either within the internal bodily milieu, by ANS reaction, or brought about by the organism's motor behavior enacted within the environment. It is crucial for the brain to infer (i.e., to anticipate) allostatic needs, in order to engage in effective action selection and thus avoid harm (Stephan et al., 2016). For our purposes it is important to note that perturbations to allostasis can occur not only from the physical but also from the social environment of the individual.

BOX 2 | The subjective experience of emotion.

Drawing on a wide range of research, Craig concludes that the convergent evidence "*implies directly that the AIC supports awareness of the immediate moment with a coherent representation of 'my feelings' about 'that thing'*" (Craig, 2009b, p. 65).

It has further been argued that homeostatic processes and interoceptive signals underpin the experience of self. The posterior to anterior re-representation of interoceptive sensation within the insula allows for the integration of the body's homeostatic condition with exteroceptive sensory input, as well as information about the individual's motivational and social context from other brain regions such as the ACC, hypothalamus, amygdala, ventral medial and dorsolateral prefrontal cortex and the ventral striatum (Craig, 2009a) (**Figure 2**). Such a perspective is supported by Critchley and Seth's conclusion that the "*insular cortex supports a neural representation of changes in internal arousal states, and, within anterior insular cortex the re-representation of this information is proposed to underlie subjective emotional feelings and their abstraction to both the encoding of future risk and the experience of empathic feeling for others.*" (Critchley and Seth, 2012, p. 424).

Damasio (2010) similarly proposed that interoception plays a crucial role in generating a subjective sense of self. However, while Craig places the self firmly in the insula, Damasio cited a patient whose anterior insula was destroyed by a brain lesion but continued to exhibit all signs associated with feelings and awareness of self (Damasio et al., 2013). A possible reconciliation of these opposing views is that the neural substrate of feeling states is to be found first subcortically and then secondarily elaborated at a cortical level (e.g., in AIC and ACC) (Damasio et al., 2013; Solms, 2019).

Building on all this, an overarching model of how interoceptive processes produce subjective awareness has been presented by Craig. The foundation of this model is in the perception of neural interoceptive signals as sensations (Craig, 2010). These signals generate pain, temperature, itch, hunger, thirst, muscle burn or ache, joint ache, sensual touch, flush, visceral urgency, nausea, among other sensations (Craig, 2008). At any given moment, the pleasant or unpleasant quality of such interoceptive sensations imbues them with motivation for the individual to move toward or away from the source of the sensation, consciously or not, while causing reactive responses in the ANS (Craig, 2008, 2010). Craig defines this functional combination of interoceptive feelings and motivation, with autonomic sequelae, as "*homeostatic emotions*," and likens them to Damasio's "*background emotions*," which Barrett calls "*core affect*" (Craig, 2008 citing Damasio, 1994; Russell and Barrett, 1999; Barrett et al., 2004).

Background emotions may be discerned through body posture, movement of the limbs, speed of motions, and animation of the face. One might use words such as "tense," "edgy," "discouraged" or "enthusiastic" as signifiers of such experience (Damasio, 1994). A similar approach is taken by Barrett, who contends that there is likely to be a core affective system which has the basic function of integrating sensation from the external world with interoceptive information. This integration generates "*a mental state that can be used to safely navigate the world, by predicting reward and threat, friend and foe*" (Barrett, 2011, p. 364).

Craig, Damasio and Barrett thus all propose that the underlying emotional experience within an individual is determined by the homeostatic management of their body's physiology, influenced by the motivational state of the body with respect to these interoceptive sensations. The whole process is constantly engaged in reconciling the past with the present moment, on a physiological level within the individual's body.

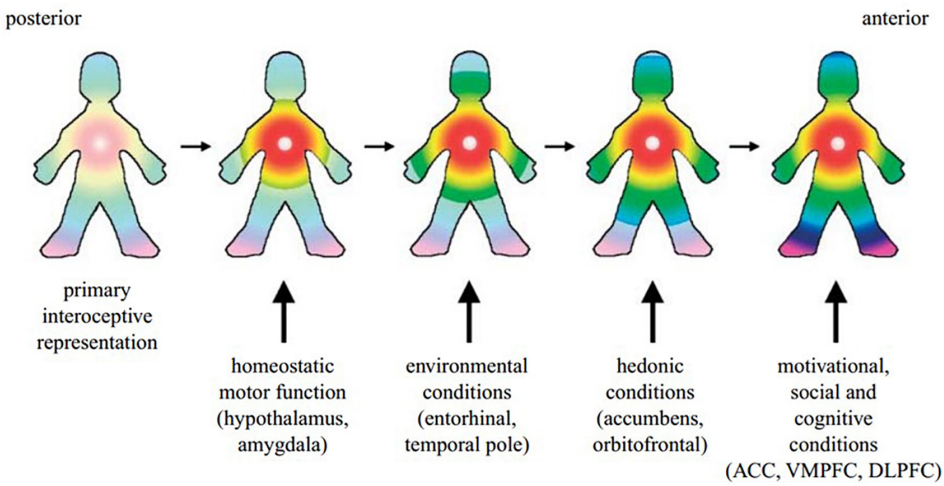

FIGURE 2 | A cartoon illustrating how the hypothesized integration of salient activity progresses from the posterior insula (left) through the mid-insula to the anterior insula (right). The primary interoceptive representations of the distinct feelings from the body in the dorsal posterior insula provide a somatotopic foundation and a template for the construction of all feelings. It is anchored by the homeostatic effects of each feeling on cardiorespiratory function, as indicated by the focus of the colors in the chest. The salient homeostatic, environmental, hedonic, motivational, social and cognitive factors are progressively integrated by the indicated inputs. VMPFC, ventromedial prefrontal cortex; DLPFC, dorsolateral prefrontal cortex. Reprinted with permission from Craig (2009a), Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1525), 1933–1942.

INTEROCEPTION, EMOTION, AND SUBJECTIVE EXPERIENCE

Current theories generally accept that emotion is embodied (Frijda, 1986, 2007; Damasio, 1994, 2010; Panskepp, 1998; Barrett et al., 2004; Wiens, 2005) and that interoceptive sensation is directly related to the functional purpose of emotion. Emotion can thus be said to be: motivation which maps the rewarding/punishing aspects of stimuli to the action system for approach/withdrawal (Rolls, 1999); necessary to compute what a stimulus means to the individual (LeDoux, 2002); and "*change in action readiness to maintain or change one's relationship to an object or event*" (Frijda, 2007, p. 158). It is interoceptive sensation itself that comprises the signal, from the body to the brain, of motivational state, with ANS sequelae as an ultimate effector arm of this processing (Craig, 2015).

What is involved in the subjective experience of embodied emotional states continues to be a subject of intense theoretical debate and this has high relevance for therapy. Two prominent early theories – the James-Lange theory (James, 1890) and that of Schachter and Singer (1962) – have claimed that we experience emotion as a result of cognitions that evaluate the changes that we perceive in the state of our body. Notably, Schachter and Singer (1962) argued that emotion involves the top-down contextualization of bodily experience by past expectations or current environment. "Bodily experience" in both these theories is what is now referred to as interoception.

In support of this, there is now substantial neuroscientific evidence that the subjective experience of emotion is generated from the integration of interoceptive signals with other sensory input, as well as top-down influences (Paulus and Stein, 2006; Harrison et al., 2010; Zaki et al., 2012; Gu et al., 2013a; Adolfi et al., 2017). Moreover, the *subjective awareness* of emotion is similarly based on interoception (Cameron, 2001; Critchley et al., 2004; Craig, 2009b; Singer et al., 2009; Damasio, 2010; Paulus and Stein, 2010; Berntson et al., 2011; Seth et al., 2011; Critchley and Nagai, 2012; Gu et al., 2012, 2013a; Herbert and Pollatos, 2012; Jones et al., 2015) (see **Box 2**).

Interoception is, similarly, vitally involved in: self-awareness (Critchley et al., 2004; Craig, 2009b); feelings of conscious presence (Seth et al., 2011); the integration of cognition with emotion (Berntson et al., 2011; Gu et al., 2013b); and empathy (Singer et al., 2009; Gu et al., 2012, 2013a). Abnormalities in the perception of interoception have been linked to: anxiety (Paulus and Stein, 2010; Duquette, 2017); depression (Harshaw, 2015; Duquette, 2017); alexithymia (Herbert and Pollatos, 2012); eating disorders (Eshkevari et al., 2014); and depersonalization (Sedeno et al., 2014; Owens et al., 2015).

PREDICTIVE PROCESSING, INFERENCE AND BAYES' THEOREM

If we assume that the integration of interoceptive sensations with exteroception and top-down information underpins subjective experience, the all-important question for therapists is how does this process produce and maintain our patients' views of themselves and the world? This question is best answered within predictive processing (Clark, 2016) where subjective experience is assumed to flow from top-down inferences that contextualize bottom up (bodily/interoceptive) sensations, while not assuming any sharp distinction between cognitive or non-cognitive processing (Seth and Friston, 2016).

As the brain does not have direct access to the world, it can only make sense of the individual's internal and external environment (or world), by inferring the causes of sensation.[2] The most interesting aspect of predictive processing theories is that they stress that our experience is largely dependent on what our brains "predict" or "expect" is happening, at any given moment, based (at least in part) on previous experience (Gu et al., 2013a; Seth, 2013; Barrett and Simmons, 2015; Seth and Friston, 2016; Ondobaka et al., 2017; Miller and Clark, 2018). Such expectations or beliefs are not necessarily conscious, nor available to awareness, for example, the "beliefs" we formed in infancy about comforting or fearful stimuli.

The predictive process that the brain uses to infer the causes of its sensory states can be described statistically by Bayes' Theorem, which provides a principled way of describing how we test and update our hypotheses (also sometimes referred to as our implicit "priors," "predictions," "expectations" or "beliefs") against the evidence supplied by our (interoceptive and exteroceptive) sensory organs, learning (and continually updating) the mostly likely cause for some particular incoming (set of) sensory information.[3] As a "Bayesian observer," the brain attempts to "know" about inner and outer experience by calling up a previously determined prediction (A) (also "prior") that seems most likely to explain the current sensory evidence and then comparing that prediction with the actual incoming flow of interoceptive information (the observation X). If they don't match, a "prediction error" occurs that may then be used to update the prior belief, creating a new ("posterior") belief.[3]

This process of evaluating incoming sensation against prior belief is hypothesized to occur throughout the brain, in a hierarchical manner, i.e., through a step-wise process of activity, within neuroanatomic (hierarchical) levels of the brain. For ease of understanding, the terminology used: "lower" vs. "higher" in reference to this hierarchy, refers to areas that are more sensory bound as "lower," while those that are more bound to prior beliefs or associative processes are labeled "higher."

Predictive processing is well understood in vision (Rao and Ballard, 1999). For example, I am walking in the street and see a shape in my peripheral vision. The sensory input to my retina quickly matches a pattern that generates a hypothesis (belief) that it is a tiger. This is my "prior." I look more carefully and see

[2]Predictive processing describes how this inferential activity may be organized within the hierarchical systems in the brain to produce generative models that are then used to infer (to generate) the causes of incoming sensory data (i.e., to explain it), whatever the form that data takes – whether interoceptive or exteroceptive Friston (2010).

[3]Frith (2007) neatly describes Bayes' Theorem as: "*Given some phenomenon (A) that we want to know about, and an observation (X) that is evidence relating to A, Bayes' Theorem tells how much we should update our knowledge of A, given the new evidence X.*"

either that yes, indeed it is a tiger (hence no prediction error). Or perhaps I see that it is a large cat! This evidence (very different from my initial prediction) results in a large prediction error and a considerable revision of my original (mistaken) prior belief about a tiger – into the posterior belief that provides a much better, error minimizing, explanation for my sensations – namely, that what I am seeing is a cat.

FREE ENERGY AND ENTROPY

As a prominent advocate and extender of predictive processing, Karl Friston's innovative thinking culminates in the "free energy principle" (Friston et al., 2006; Friston, 2010, 2013) which asserts that in order to maintain homeostasis and survive all living organisms must avoid surprising states (i.e., free energy). Clearly, when applied to interoception, minimizing surprise is just another way of describing homeostasis (i.e., keeping interoceptive signals within a comfortable and familiar range). At a more general level, human beings don't like surprise – or its mathematical average – namely, uncertainty (Edwards et al., 2012), which is equivalent to summed prediction error, i.e., the difference between what our brains predict and what our actual sensations are at any given point. Minimizing free energy is equivalent to resisting entropy (the tendency for a system to become disordered, dissipate decay and ultimately die). The beauty of the free energy principle is that it accounts mathematically for the inherent drives that biological organisms have toward allostasis and self-organization.

In effect, free energy can be thought of as the difference between the actual states of an organism and the states it "believes" (generally unconsciously) that it needs to be in for its adaptation, survival and reproductive success. When a prior belief doesn't match the incoming sensory data, there is prediction error (i.e., uncertainty or free energy), which we can reduce by updating our beliefs (from a prior to an improved posterior belief) about the state of affairs in the world. In other words, updating our priors makes our predictions better explanations for sensation and thus minimizes prediction error. This process of updating priors is "perceptual inference." Assuming that priors are generally encoded in higher levels of the brain's hierarchies, this implies that prediction errors ascend the brain's hierarchy to do the updating. However, all animals can instead act to change the world (and their own body) to make the sensations that they receive fit with their predictions. For this to happen, at the very bottom of the hierarchy (i.e., at the most sensory bound level) prediction errors descend to activate the effector organs to elicit motor reflexes. This is known as "active inference." In our example of the supposed tiger in the street, the original belief that there is a tiger immediately invoked a higher level (learned) cognitive prior that tigers are not generally seen in suburbia. To resolve uncertainty about what is actually causing visual impressions, the brain predicts that it will foveate the supposed "Tiger." These predictions about the (proprioceptive and exteroceptive) consequences of "looking over there" are then issued to the oculomotor system. In turn, the oculomotor system resolves proprioceptive prediction errors by using them to drive motor reflexes that point the eye to the predicted (i.e., intended) location. In essence, this is active inference, where prediction errors drive bodily changes to eliminate themselves and – in so doing – fulfill top-down predictions.

Predictions or beliefs can be straightforward, e.g., about how far our hand must move to reach a pen, or they may be highly abstract and refer to the social intentions of another human being. Some predictions will be innate and not subject to updating (such as homeostatic set points) but others are at least partly learned – many in infancy. Updateable priors include priors about policies/actions (e.g., priors about using a habit); priors about models of the world, and priors about beliefs within a generative model of the world. Importantly, some priors must govern the supposed reliability and salience of any given incoming sensation (these are priors about "precision" – discussed below).

If there is discordance between sensations and the expected cause of those sensations, surprise, uncertainty and free energy are ostensibly increased. Minimizing prediction error/free energy from moment to moment ensures that human organisms adapt and survive. In a perfect world, free energy minimization thus results in healthy, optimal functioning. However, the world isn't perfect – it is predictably unpredictable.

Combining the biological imperative that humans are at equilibrium with their environment (internal and external) when free energy is minimized, together with the psychological assertion that human beings avoid pain and approach pleasure, Joffily and Coricelli elegantly link free energy, or uncertainty, to the valence of the emotional state. They suggest that if there is a rise in free energy, due to a mismatch between expectations and sensory input, there will be more inherent surprise or uncertainty, producing an emotion with negative valence. Examples are fear, disappointment and unhappiness. If there are prediction errors but the opposite is true and free energy is falling, then the valence of the resulting emotion is positive, such as for happiness or hope. Intriguingly, Joffily and Coricelli go further and suggest that fear can be distinguished from unhappiness because fear involves not only rising free energy (the organism is moving away from desired set points) but also includes the expectation that this rate of change will accelerate. Unhappiness, by contrast, involves only that free energy has risen but with no expectation that this change will get worse. For happiness the expectation is that the fall in free energy will accelerate, while for hope there is no such expectation (Joffily and Coricelli, 2013).

ACTIVE INFERENCE IN THE INTEROCEPTIVE DOMAIN

As outlined above, prediction errors may update perceptual priors (this is a case of changing the brain's inner model to better fit the world). Alternatively, they may be resolved by descending to the brain stem and driving motor reflexes so that the animal acts on the world, which may serve to make the world a better fit for the prediction.

In the case of interoceptive prediction errors, descending prediction errors can enslave ANS reflexes (e.g., by raising heart rate in response to the perceived threat). Thus, in our tiger example, a high-level cognitive prior (learned or partly innate)

is also immediately invoked that the viewer is in danger. This prior sets up prediction errors between the currently relaxed state of the body and the state it needs to be in to evade a predator. These descend to activate ANS reflexes and are eliminated by invoking high arousal. Prediction error is thus minimized by changing the body to better fit the world (Critchley and Seth, 2012; Gu et al., 2013a; Seth, 2013; Hyett et al., 2015; Seth and Friston, 2016; Critchley and Garfinkel, 2017). This process, known as "active inference," is a special case of prediction error (or free energy) minimizing in the interoceptive domain; where action corresponds to autonomic regulation (Critchley and Seth, 2012; Seth, 2013; Pezzulo, 2014; Barrett and Simmons, 2015; Seth and Friston, 2016; Quadt et al., 2018). This circular signaling between body and brain, causing autonomic reactions, can serve as an important entry point into therapeutic change, as we discuss below.

The literature on reinforcement learning and the formation of "habits" is relevant here. It is proposed that reinforcement learning takes place in two ways. Model-based (goal-directed) earning is essentially Bayesian, whereby the learner has a model (e.g., a prior belief) that updates in the light of available evidence. In model-free (habitual) learning, on the other hand, "*through the process of sequence learning, action control becomes increasingly dependent on the history of previous actions and independent of environmental stimuli, to the point that, given some triggering event, the whole sequence of actions is expressed as an integrated unit*" or habit (Dezfouli and Balleine, 2012, p. 1038). Many empirical studies (in rodents and humans) show that, during reinforcement learning, behavior is initially goal-directed but then becomes habit-based. Importantly, it has been shown empirically that habits are insensitive to both changes in the context and changes in reinforcement (Dolan and Dayan, 2013). In other words, they persist inappropriately. In Friston's words "*after the habit has been acquired, there is no opportunity for pragmatic policies. This means that although the behaviour is efficient in terms of reaction times, the habit has precluded exploitative behaviour*" (Friston et al., 2016, p. 874). In this way our patients' emotional states can reinforce a style of interacting with the world (habits) on the basis of active inference processes gone awry and not in their best interest.

As a keen observer of the patient's experience, the therapist is poised to support the patient in affecting change in the role that dysfunctional priors (all too often from model-free/habitual learning) play in their feeling experience, bodily reactivity, behavior and thinking processes. The therapist, alert to when the patient's bodily reaction may be patterned along beliefs other than those realistically related to the present moment, can be the initial observer of these and invite the patient to actively attend to the sensory experiences of their body. Attention invokes the crucial role of "precision" in predictive processes.

INTEROCEPTIVE INFERENCE AND SUBJECTIVE EXPERIENCE – THE ROLE OF "PRECISION"

Within predictive processing theories it is well-understood that "prediction," "prior" or "belief" do not denote a consciously held belief, but refer to activity occurring in the brain, which is assumed to encode probability distributions (i.e., subpersonal Bayesian beliefs). These distributions reflect the likelihood that a particular prior is a good explanation for the current sensory input that the brain is receiving (e.g., that "tiger" best explains the sensory pattern on my retina). There is uncertainty (variance), associated with any probability distribution. The inverse of this variance – known as "precision" – is the salience, confidence or reliability attached to a particular prior or prediction error (Friston, 2009). Within the brain, at any given moment, the attached precision (i.e., the relative reliability/salience) of the bottom-up sensation vs. the top-down prior belief is modulated by many factors (such as attention and motivation). The relative weight, i.e., the precision of the prior belief vs. the incoming sensory information is the determining factor in whether updating of a prior occurs. Prediction errors stemming from sensation that is precise (reliable) will update an imprecise prior. However, a highly precise prior (e.g., a historical prior or habit that has preverbal roots from infancy) may resist updating. For example, the prediction errors for danger invoked by the possibility of a tiger are highly precise and drive ANS as well as motor reflexes. Nevertheless, the prior for a tiger loose in a city is very imprecise (unlikely) and is easily updated to "cat."

Attention is a key driver (or psychological homolog) of precision (e.g., Feldman and Friston, 2010; Edwards et al., 2012). Pezzulo (2014) uses an engaging story about a dark night, a creaking shutter and fears of an intruder (the imaginary bogeyman), to illustrate this. He notes that interoceptive sensory information is often afforded more attention than exteroception because it is commonly experienced as more certain by the individual, thus maintaining higher precision relative to other sources of sensation and consequently asserting a disproportionate effect in the inference process. Pezzulo describes how the resulting affective experience and the resulting physiological reactions might create a belief of immediate danger (from a bogeyman) in the middle of the night. Importantly for our purposes, he describes how shifts in precision due to unrecognized attentional imperatives can result in experiences and behavior that seem to reflect the reality of the moment, yet are actually the result of significant distortions (Pezzulo, 2014).

In our patients' models of the world, precision dictates how strongly they hold to their priors, in spite of evidence to the contrary. Patients suffer where they rely on highly over-learned and thus very precise priors which do not reflect the truth but are difficult to update, as they have gained relative strength with repetition over time and have become habits that are insensitive to changes in context or reinforcement. For example, for an anxious person at times of fearful distress instigated by perceived threat, the default prior "danger" will be afforded higher precision. As a result, prediction errors signaling that they are actually safe don't update this prior, as the individual fails to attend to, and thus increase the precision of, disconfirming evidence in the here and now. Crucially, these prediction errors can be resolved instead by changing the body so that it fits the habitual prior (e.g., by raising heart rate to fit the prior for threat). In other words, interoceptive prediction errors instigate the very autonomic reactions that drive the ANS into inappropriate

arousal (freeze, flight, fight) in order to confirm the habitual prior. This high bodily arousal will be experienced as fear, as the emotional state updates to fit the incoming information of arousal. From the perspective of the patient's emotions, this illustrates the classic James-Lange contention that we feel fear when we receive peripheral information from the body.

What is striking here, for psychotherapeutic treatment, is that the *expected* interoceptive arousal state that corresponds to fear – such as muscle tension, heart rate increases, or hormonal response – is actually being *produced* within the body in response to the initial highly precise prior of perceived threat, although such a threat does not actually exist at that moment. Such conceptualizations explain the observations at the beginning in this paper that a patient will be certain that she is scared only because her heart is beating faster. This contention can now be seen as circular causality between the brain and the body, provoked by active inference. Interoceptive prediction errors have descended to activate ANS reflexes that affect the body by raising heart rate. Sensation that has been created by ANS activity then returns up the hierarchy to the brain and verifies fear as the person's "*value-based choice about the internal state of [their] body*" (Gu and FitzGerald, 2014, p. 269). In other words, the high precision of the habitual/default prior for danger unfortunately specifies that fear is indeed a predicted and familiar state for the individual to be in, which itself reinforces the precision of the prior rather than there being any attention to disconfirming evidence. Sadly, the brain returns the body to a state of fearful arousal, simply because this is the expected state, ostensibly with free energy lower overall as there is less uncertainty (Friston, 2010; Clark, 2016; Ondobaka et al., 2017). Such moments – when precise prior expectation cause ANS reaction and thus bring about the perceived confirmation of the original expectation – are important points of therapeutic access, which we will address in detail below.

Some priors are obviously more resistant to updating from sensory evidence than others. Yon and colleagues point out that if the brain were actually an "*ideal scientist*," as predictive processing explanations imply, then a hypothesis about the causes of sensation would simply be compared to the evidence and the brain would update this hypothesis accordingly. But in real life priors are often resistant to disconfirmation, either as a result of evolutionary pressures (as in the case of homeostatic set points) or through developmental processes, creating a situation where the brain acts more like a "*stubborn scientist*" (Yon et al., 2019, citing Bruineberg et al., 2018). This characterization highlights how important it is to investigate and address possible sources of resistant prior beliefs or habits. All psychological treatments have some process orientation that supports the instigation and testing of hypotheses by their patients.

For the purposes of this paper, we focus on infancy as a vital time for the inferential processing of experience into beliefs about self and others, habits, as well as the linking of emotion and interoceptive processes and experience, within the caretaking relationship. The psychological literature centers on the importance of attachment in how an individual will react to strong emotions (e.g., Stern, 1985; Bowlby, 1988; Siegel, 1999). For example, the phrase "*implicit relational knowing*," used by Lyons-Ruth et al. (1998), Morgan et al. (1998), and Stern et al. (1998), places the development of procedural knowledge concerning both interactive processes and affective experience within the relational interactions between caretaker and infant. Such caregiver-child interactions are especially important with respect to periods of strong arousal, which is considered to be the burgeoning experiential element of early emotions (Schore, 2003).

It has been proposed that physical contact with caretakers in infancy acts as an early homeostatic regulator, supporting the development of the immature nervous system (Fotopoulou and Tsakiris, 2017). Furthermore, it is argued that through the quality of caregivers' understanding of our body's needs, coincident with our nascent inferential processing of interactions with them and with our inner and outer environments, our brains develop early (unconscious) "prior beliefs" about the causes of our sensory states. A child's brain, for example, forms beliefs about what situations are comforting or fearful through caregivers' effective (or ineffective) provision of necessary resources. Such preverbal experiences (with subcortical representations that may be more salient – i.e., have higher precision – than any later cortical elaboration) strongly influence our physiological regulatory processes and become imbued with motivational significance, thus forming the basis of our subjective emotional experience and our interactions in the world. Activities within an attachment relationship that settle the infant's nervous system occur following successful interactions between the infant and the primary caretaker. In this case, interoceptive signals facilitating homeostatic balance within the emerging predictive system of the infant then support a sense of physiologic stability and overall well-being (Fotopoulou and Tsakiris, 2017).

While addressing the importance of physical contact and interpersonal interactions with caregivers on the development of the self, Fotopoulou and Tsakiris (2017) link such bodily based interactions with the development of early "*mentalization of interoception*." Importantly they note that it is the direct proximity, style of contact and care of the infant's homeostatic needs which are important progenitors of experience. As such interactions become more complex, the growing child can "*build increasingly more sophisticated models of their own interoceptive states, as well as strategies for minimizing free energy in the interoceptive system*" (Fotopoulou and Tsakiris, 2017, p. 17).

Commenting on the originally subcortical nature of emotion (which is subsequently elaborated in the cortex), Solms (2019, p. 13) makes the important point that "*subcortical memory traces cannot be retrieved in the form of images for the simple reason that they do not consist in cortical mappings of the sensory-motor surface organs*." This has the crucial implication that feelings states (emotions) that have been acquired as habit sequences in infancy may lack cortical expression. It is our contention that these may, therefore, only be accessible through the bodily (interoceptive) representations that accompany them. Daw has suggested that "*hierarchical reinforcement learning decompose a multistep decision problem into a nested set of choices at different levels of temporal abstraction*." He proposes that lower-level choices are essentially "*model-free: stereotyped behavioral*

sequences, like a tennis serve or a dance move" (Daw, 2015, p. 13750). If this is so, then our patients' precise priors (that appear unavailable to updating) may be characterized as low-level habits or routines to which the patient has little conscious access, other than that they create interoceptive (ANS) sensations. It is therefore by attending to these interoceptive sensations, in the safety of the therapeutic relationship, and by challenging their relevance to the here and now, that we can increase the precision of disconfirming evidence and hopefully move the patient from a model-free reaction to one that is model-based, in which current evidence can be evaluated and can update the model.

IMPLICATIONS FOR PSYCHOTHERAPY

As discussed above, it is generally accepted that emotion is always accompanied by interoceptive sensations within the body that signal what is motivationally salient to the individual. These sensations can become conscious "feelings," which we define here as emotions that are known and/or verbalized. While not all emotion reaches awareness in the form of feeling, each psychological school of thought has cogent theoretical reasons for why it is therapeutic for the patient to bring emotions into subjective awareness. Our focus here is to propose that attending to interoception is crucial in this context.

Emotion, defined as sensation with coincident motivation and resulting autonomic sequelae (Strigo and Craig, 2016, citing Rolls, 1999) can be described through the processing of incoming interoceptive sensation that is integrated with concurrent exteroceptive sensation (Seth and Friston, 2016). The process by which this occurs, i.e., the "generative model" that is used in the current moment (with its priors and precisions at every level of the hierarchy), will tend to follow a pattern (i.e., a habit), that often represents a view of the world and the self that was experienced within the early caretaking relationship. Such reactions are not consciously remembered *per se*, but exist as a set of bodily felt sensations that persist due to the salience (the precision) that was afforded to the various sensations in these early relationships. For example, if the infant was not responded to with consistent signaling of safety/certainty by caregivers – communicated at the bodily level – the prior for threat will be highly precise. That adult individual will tend to stay in a state of increasing uncertainty about how others will respond and may be unable to take account of disconfirming evidence. There follows the crucial insight that this uncertainty/anxiety may actually be a highly familiar state to the individual. The pernicious potential effect is that the brain will tend to accept this state as that to which it should seek to return the body, in order to minimize free energy, despite the fact that this state is represented by the patient as distressing. For psychotherapeutic treatment, the implications of interoceptive inference are thus profound. Our task is to revise such familiar, but dysfunctional, priors (habits) to which the patient tends to return. However, the power (i.e., precision) of such priors highlights the early experiences of the body, which may not have an explicit conceptual component although they have bodily/emotional precisions that are very important for determining the individual's current state. Our proposal is that acknowledging the influence of bodily sensation on the processes of mind, and creating a window into the influence of the body and sensory experience on emotional states and cognitions, allows therapists a much wider range of interventions based on the body than when engaging relationally with our patients in "just" talk therapy.

To illustrate this, it is important to remember that within a predictive processing view emotions and feelings are always hypotheses that provide the best explanation for the many interoceptive and exteroceptive cues which have to be explained. For example, the (high-level, conscious) hypothesis "I am anxious" may be the best available explanation for a breadth of visual, auditory, somatosensory and (crucially) autonomic sensations that are currently in play. Therapists can recognize (rather readily) that there may be several other possible explanations, yet patients are likely to reference a habitual explanation for the stimuli that are commonly salient to them, often creating a "story" regarding their meaning, even if such responses are sub-optimal or create dysfunction. In determining the point at which an intervention can be implemented therapeutically, we make an important distinction between: (i) "active interoceptive inference" in terms of a pre-reflective, subpersonal assimilation of interoceptive cues that contextualizes our autonomic reflexes and reactions to emotionally salient cues; and (ii) the personal, reflective or propositional inference involved in "mentalizing interoception," that raises emotions into subjective awareness (as feelings) and thus involves explicating the content of active (interoceptive) inference that is available to conscious awareness (Fotopoulou and Tsakiris, 2017; Ondobaka et al., 2017).

To reiterate, bodily-based experiences will often over-determine current perceptual content (i.e., emotional priors). There may be a habitual and sub-optimal prior (generally unconscious) that connects some harmless stimulus to a sub-personal interoceptive inference, such that the body responds as if there is a threat, with ANS sequelae. At this juncture, however, if the patient is to engage in the selection of the emotional state that is a better explanation for the sensory evidence at hand, and reality in the present moment, s/he has to have more than one hypothesis (prior) available. In other words, s/he has to be able to select one emotional hypothesis over another, particularly in times of emotional distress, in order, for example, to distinguish "I am just excited" from "I am anxious." The therapist cannot draw attention to the low-level prior that set the bodily/emotional reaction into play (because this is subpersonal, or unconscious). However, by drawing attention to the patient's physical reaction (the ANS sequelae) – which is available to awareness – they can heighten the precision of the interoceptive evidence, thereby potentially allowing the inferential process to affect the habitual prior. This creates an opportunity to explore new priors, i.e., alternative explanations for these (now attended) bodily sensations, thus enabling interoception to be consciously mentalized. Achieving this desirable "mentalizing of interoception" necessarily requires elevating (low-level) subpersonal emotional experience to a higher, reflective level that, in turn, requires attending to pre-reflective interoceptive inference i.e., attending to current

interoceptive sensations (Barrett et al., 2016; Hoemann et al., 2017). This mentalizing ability can, however, only emerge if we first attend to (and thus increase the precision of) interoceptive sensations.

Given that predictive processing theories indicate that to alter habitual ANS reflexes requires a change in the precision of the sensory evidence, to achieve this within the therapeutic interaction, requires the patient's active attention to current bodily sensations, for example, the sensations associated with their body in the chair and the feeling of the ground beneath their feet. The mentalizing of interoception ostensibly starts as an "imaginative" process of exploring bodily sensation. While, strictly speaking, this also involves somatosensation, if the patient attends and tries to receive, assess, and appraise the embodied nature of all the sensation perceived, the result will necessarily involve interoceptive appraisal (Farb et al., 2015).

Focused attention increases the precision of current bodily sensations (thus generating ascending interoceptive prediction errors). There is now increasing uncertainty. This at first leads to an increase in the intensity of the experienced feeling, which the patient will find unpleasant and to which the therapist must consequently respond. The therapists' persistent observation and enquiry into the patient's experienced state of their body is vital at this juncture. Observing physical expressions such as eye contact, facial and body movements, prosody of speech, muscle tension, will facilitate a qualitative assessment of changes in the state of the ANS (Schore, 2012).

Furthermore, the therapist must remain alert to many other variables in the therapeutic process which affect emotional valence and which can cause change from positive to negative valence (representing dynamic changes in free energy), without the patient necessarily being aware. For example, sudden changes in the manner in which patients express feelings verbally (less spontaneous verbalizations, word choice, prosody), or alterations in the patient's physical response (more body tension, sitting forward/back, folding up their legs and arms) could indicate changing emotional valence within the patient and rising or falling free energy. Pointing out bodily changes to the patient, and encouraging them to reflect on what the therapist observes, can bring to the fore contextual aspects of the present moment that are stimulating increasing uncertainty in the patient but were unrecognized by her. We propose that these physical expressions are linked, ultimately, to early beliefs and relational expectations, which are not recognized cognitively but are being expressed bodily. It is our contention that when these come to awareness in the patient, this provides an important element that supports change, in any therapeutic process.

Uncertainty is constantly being evaluated across multiple hierarchical levels in the brain at any moment, influencing experience and behavior. As the therapist models recognition of bodily changes to the patient and enquires about emotional valence, and the change of valence in either direction, the patient can better take a reflective position regarding their bodily-based responses to their internal and external environment and the attached uncertainty. This is a functional benefit of mentalizing interoception. A unique blend of possible insight is available to the patient at such times – into affective experience, cognitions and current moment perspective – which also invite a process of change which the patient can implement when they are no longer in therapeutic session.

As the patient tries to discern the inherent faults of their models of the world, the therapeutic relationship should offer a "safe haven," otherwise, a sense of threat may irreparably impact the freedom the patient requires to explore. The safety of the therapeutic relationship is paramount and must be attended to, explicitly and implicitly, between therapist and patient. The therapeutic attachment relationship in widely thought to be important in supporting not only the perception of safety but also to facilitate change (e.g., Cozolino, 2002; Wallin, 2007; Siegel, 2010). While a detailed discussion is beyond the scope of this paper, we will focus on an element of the therapeutic relationship, "the real relationship," which has specific aspects that can be operationalized regarding the therapeutic relationship, the treatment environment, and the predictive process aspects of the mentalizing of interoception. The real relationship, is defined as the non-transference part of the relationship (Gelso, 2009), concerning the authentic, genuine and realistic aspects of the relationship between therapist and patient (Duquette, 2010). We briefly outline below what the real relationship brings to the process.

The consistent presence of the therapist, experienced repeatedly in times of uncertainty creates a history of moving together within the therapist/patient dyad. And hopefully, such times resolve for the patient with an increasing sense of verifiable safety amidst open vulnerability. Such history "*encourages the patient to move 'deeper' into the chaos experienced at any given moment*" (Duquette, 2010, p. 141). The experienced authentic, reality based, qualities of the relationship also presents a space that is distinctly different, and not laden with the expectancies that occur with people the patient knows (Morgan et al., 1998). Ultimately, as such elements of the relationship lessen uncertainty in the therapeutic process, thus decreasing prediction error, persistently supporting a gradient of decreasing free energy.

All schools of psychotherapeutic treatment must address their patients' persistent prior beliefs, that are manifested in the form of cognitions or behaviors that stem from inaccurate priors that persist in the face of a different current reality. Therapists use every available resource to encourage the patient to become a flexible scientist and test their strongly held hypotheses against a different current reality. How the therapist will choose to support the shifting of priors, using the direction of attention, will depend upon the form of relational interaction delineated by their particular protocol. For example, in Acceptance and Commitment Therapy (ACT) the identification of old beliefs about the self will be paramount, together with means to direct attention away from these toward disconfirming evidence. In cognitive behavioral therapy (CBT) the identification of cognitions is key, with attention being directed to unlinking the associated reactive behaviors. The common experience of transference can also be interpreted within a predictive framework – as a process that molds the patient's past priors onto the present relationship – while counter-transference is the functional equivalent for the therapist. The various therapeutic

schools will view the outcome of predictive processes (including transference reactions) through the lens of their own philosophy but all therapists will be better equipped if they recognize the inherent difficulty that the patient has in trying to be an impartial Bayesian scientist with respect to the sensory data of their inner and outer world.

DISCUSSION OF CLINICAL VIGNETTE

The material and concepts outlined above inform the case study contained in **Box 3**. [Written informed consent was obtained from all mentioned individuals for the publication of this case report/case description, no identifiable information is included and pseudonyms were used]. Elements of Molly's clinical interactions with the therapist illustrate how the neuroscientific ideas described above are useful in psychotherapeutic treatment.

Within Molly's history there are important clues from her early life that foreshadow her difficulties with physiologic and emotional regulation. Her mother is a highly inconsistent presence during her first year and likely longer, as Molly describes that her family viewed her as independent from an early age. There is evidence of homeostatic disruption: her sense that "something bad was going to happen" as she awoke; her body either over-reactive (moving all the time) or under-reactive (sudden bursts of tears); and her strong sense of discomfort in her body; and that her needs will be seen as an imposition on others. Such symptoms echo the premise that homeostatic processes are "*dependent on embodied interactions with other bodies*" (Fotopoulou and Tsakiris, 2017, p. 13).

The initial statement made by Molly, "*Things just had to be different, life shouldn't be this hard*," is pertinent. Habitual behaviors are leading to poor outcomes as she enters adulthood, and she can't generate alternative hypotheses. She recognizes that she can't successfully problem-solve even the minor issues in her life and reports her view of herself as "*inadequate and needy*," which causes problems in interactions with others. Such persistent negative expectations create a snowball effect within Bayesian processing for patients with depressive disorders (Barrett et al., 2016). Such patients are not able to appraise disconfirming evidence adequately, either by discounting its credibility, or by seeing it as the exception rather than the rule, Molly's ruminations presenting evidence of only her original hypotheses and beliefs (Rief et al., 2015; Kube et al., 2017, 2019).

Considering mentalization from the perspective of intentional mental states (Bateman and Fonagy, 2012), Molly's initial inability to use language to express her emotions – only being able to know that life "*should*" be different, and her fear or despair is "*something bad*" that might happen – is indicative of mentalization in that form. The possible importance of her mother's absence due to hospitalization and depressive symptoms, is indicated by Fotopoulou and Tsakiris (2017, p. 17) claim that "*the origins of all mentalization processes are not only embodied but also by necessity involve other people's bodies, their physical presence, proximity, contact and most importantly, their homeostatically relevant actions*." It appears that Molly's temperament, physicality and the actions of other caretakers may have helped her push through such early deprivation. However, ultimately the effect of the impaired regulatory processes that she would have experienced with her mother's long absences in her infancy are evidenced in her symptoms in each domain.

Early in the session, as the therapist enquires about Molly's constant physical movement. She is initially unable to be still or to respond in a reflective manner and she expresses a strong sense of threat. Stillness would only leave her in "*an empty place*," "*painfully uncomfortable for me*." The interceptive sensation from her body simply reinforces Molly's prior beliefs and fearful state. She cannot intentionally pursue any deliberation. At this juncture the therapist offers an alternative view. She encourages Molly to imagine whether she feels the pain inside or outside her body. When Molly begins to take part in this cooperative narrative about her discomfort, the therapist puts herself forward as a safe space in which to try out something new by inviting Molly to test her negative expectation "*here with me*." The therapist encourages her to look for new evidence by making eye contact with her, while stressing "*the obvious elements of safety of the moment*." In several ways the therapist thus scaffolds Molly's efforts to mentalize her interoceptive experience, ultimately resulting in a noticeable difference in the activation of Molly's body, with deeper breathing and tears.

Molly's reactions to the therapist's interventions relate directly to how the body can be supported within relational interactions, allocating resources more effectively allostatically, thus effectively altering the ANS reactivity that stems from suboptimal habitual priors. The physical changes observed – the slowing of her breathing, increased eye contact and lessened bodily tension, evidence a shift in Molly's autonomic state from a high sympathetically driven state toward a state with more parasympathetic control (Zautra et al., 2010: Vlemincx et al., 2015; Strigo and Craig, 2016) and indicate an updating of priors.

Attending to sensory experience from a position of observation, without judgment, allows the individual's higher-order cognitive processes to shift into state of more open consideration and observation (Farb et al., 2015). This permits a new flexibility, which can facilitate awareness of interoceptive sensations, to which the patient may not habitually attend, and which may promote positive experiences and lessen the individual's automatic return to cognitive elaboration or "stories" (Fogel, 2009; Farb et al., 2015). Ultimately, there will be a decrease in the general energy output (i.e., free energy) that a person may typically spend on self-regulation, especially if they employ active inferential processes (i.e., firing up ANS reactivity), in which they are effectively trying to change their own bodies to fit the dysfunctional habitual prior to which the brain continually seeks to return them. This can be seen in Molly's original assertion that "*I have to keep moving to be safe*" which finally shifts to less tension. Deliberate intention is required to move into such a more observational mode with respect to experience. Initially such a relaxation of energy may feel difficult for the patient, due to the automatic nature of the habitual prior. However, with practice, the sense of relief that can follow a change from lower to higher free energy (Joffily and Coricelli, 2013) can positively encourage the patient. Notably, Molly's activity level did not increase again throughout the remained of her first session.

BOX 3 | Clinical case study.

[Written informed consent for the publication of this case study was obtained from all the individuals mentioned. No identifiable information is included. Pseudonyms are used throughout.]

Molly was 20 years old when she entered long term individual and group therapy. She made an appointment because her sibling who was in treatment suggested she do so, "If you need someone to talk to." Molly could only express a sense that somehow "things just had to be different, life shouldn't be this hard." When Molly was a year old, her mother was psychiatrically hospitalized for major depression for several weeks and was thereafter often debilitated. Molly was considered bright and independent by her family, which was important to her sense of herself, an example she remembered was that by the age of seven she would travel miles away from their suburban home on her bicycle, only returning in the evening. At presentation she could describe that she was readily upset by small problems, had difficulty in relationships as she now saw herself as "inadequate and needy," feeling constant self-criticism about her interactions with others. She complained of often not falling asleep due to ruminative thinking and having the sense "like something bad was going to happen" as she awoke each morning.

She rarely sat still, a leg jiggling, shifting in her seat, moving her hands. If asked what she felt in her body, she looked perplexed and said, "Nothing." Although she expressed little insight into her emotional or physical state, she was quick to tears when touched by strong feelings, and would tense her facial muscles and throat to limit their expression. Her answers were often content-based and lacked contextual depth. However, with a clear sense of empathy for other's experiences, Molly rarely was critical of others, but of herself in most instances.

In an individual session early in her therapy, Molly's constant motion was addressed by her therapist. "You seem to be constantly on the move, any feelings that you are aware of?" She recognized she did "move a bit but if I don't I will be bored." When asked what "bored" meant, she could only say it was like, "an empty place, that just isn't good to be in." Her therapist then asked Molly if she could sit entirely still and Molly immediately flatly insisted, "No that would just be too uncomfortable." Quizzically the therapist asked, "Uncomfortable? Sitting still is more uncomfortable than moving your legs, hands, and shifting in your chair so often?" "Sitting still is just uncomfortable, like painful uncomfortable for me, maybe not you, but for me it is." Does it feel as if the painful sense is in your body somewhere, or does it feel as if something painful will come from outside your body if you sit still?" Molly paused for a minute, "Hmmm, well now that you put it that way. . .I know I do feel uncomfortable in my body, I don't know but maybe it's that thing you call feelings? I just never sit still with people and it is just safer all around."

Her therapist offered, "How about you give it a try here, with me. How about you put your feet on the floor and your hands in your lap, just for 30 seconds now?" Molly resisted through a few rounds of interaction, then skeptically placed her feet on the ground, and put her hands in her lap. She didn't make eye contact and she held her body rigidly in place, but her breathing slowed gradually as her eyes softened and began to well with tears. "Any idea of what your tears are about?" her therapist asked. After a pause, Molly said, "It just feels like in the world there is too much that will hurt me, and I have to keep away and keep moving to be safe." Her therapist invited her to make eye contact, take a deeper breath, while also talking about the obvious elements of safety of the moment and how she, the therapist, was alert to any possible dangers for Molly. Slowly Molly raised her eyes, took a slightly deeper breath as tears fell more readily from her eyes. She said slowly, "It just feels like I have this pressure in my head, like at my forehead, like all my feelings are bound up there." Her body had become less tense, her voice softening. Her activity level didn't increase again throughout the remainder of the session, exhibiting a deepening emotional involvement with her therapist.

Molly drove an old car and she could make any necessary repairs, which were frequent. One day she repaired her car alone on a city street in a dangerous area, rather than asking anyone to help her fix it or get it towed to a safer place. She came to group directly after that repair, with dark grease on her hands and her clothes disheveled. A group member, visibly upset, said "How could you just start working on your car in such a dangerous part of the city, and alone, too!" Molly answered caustically, "And what was I supposed to do? I had no choice, I had to get here, I don't need to ask anyone for help, that's how I could work on it on the street. I don't care where the street is!" The other patient asked, "Did you even think about asking anyone for help when your car broke down?" "No, I've never asked anyone for help when I can take care of the problem." With a softer voice, the other patient said, "But you were on a street in the city alone, you could have been in a lot of danger, just having someone nearby would have been a help, wouldn't it?" Molly became more physically agitated at that interchange. "I don't care that I was alone, I can take care of myself just fine, that's just the way you do things, no one is there to help with problems, they just add to them, you just don't understand." With that Molly sat back in her chair, arms and legs crossed with a deep scowl on her face and looked at the carpet.

Her therapist spoke to her next, saying softly "Any idea of what you feel?"

"They're an idiot, that's what I feel!"

"Well, Molly, "They're an idiot" is not a feeling—any emotional feeling, hurt, scared, angry?"

"No, well, maybe pissed!"

"Hmmm, your voice has tears in it, does that fit with pissed?"

"I don't know, they can be as critical as they want, I take care of what I have to. However, I have to take care of it, and I don't need anyone's help."

"Right now you appear to need some kind of help, you have tears in your eyes and your voice, and a frown on your face, your body is folded up tight."

"I'm just pissed." Molly said this with tears welling in her eyes, her eyes focused on the rug in the room, and her body tensing more.

Molly continued to frown, while tears spilling down her face, which she wiped away with force. She was invited to unfold her arms, and to set her feet on the floor. She did this, reluctantly, her eyes remaining focused on the rug in the center of the room. The other patients looked concerned and at a loss as to how to approach Molly as she interpreted their compassion as judgment regarding her choices.

The therapist believed that Molly felt not only threatened by the other's opinion but also by the compassion inherently expressed in the group's stated concerns, which was disorienting as it was unfamiliar to her. To manage this feeling of threat, Molly had withdrawn from emotional contact with others and her body was shoring up her defensive position, both against interaction with others and self-experience. She believed the other misunderstood her independence, essentially belittling it when Molly wore it somewhat as a badge of honor. She could not access any sensation in her body, and was unable to think through any other possible explanations regarding her experience or actions and others' intentions. Her therapist recognized that Molly had implemented the only action plan she had known throughout her life, while repairing her car, and was continuing to do so in the therapy session, addressing only the content of the issue in front of her. As her therapist recognized Molly's body was highly reactive with fear, she understood that questioning Molly directly regarding her emotional state would only draw Molly's bright brain into the answer, forcing Molly to create a "story" to explain her beliefs about her choice to repair the car, other's intentions, and insist that she needed no further help from any others.

Her therapist asked, "What do you feel in your body right now?"

"Nothing. . .maybe some tightness"

"Can you say where the tightness is most?"

"All I know is I can feel that ball of tightness up in my head, here, near the front of my forehead." Molly ducked her head further into her chest, eyes downcast.

"So, there is tightness. Does it feel like this tightness is trying to keep something inside, or keep it outside?"

Molly considered the question for a while and then answered hesitantly, "Inside?"

"Why would you need to keep something inside?"

"I often have this feeling in my head, like there is something physically bound up in there. Maybe it is feelings but it sure seems physical to me. I've always felt like this if there was upset anywhere, it just takes me into myself, where I know I'm safe." Molly's voice softened a little.

(Continued)

BOX 3 | Continued

"You say it may be feelings bound up in there. So, if those feelings moved, or were felt, it wouldn't be safe for you? What could be the danger?"

"If the feelings moved others would be able to see them. And if they see them, they would react in a way that wouldn't be good. It never has been good." Molly's voice sounded less tense and reactive with this answer.

"You would be vulnerable if they saw your feelings, eh? Your voice is changing now, can you feel any other sensation now?"

"I can feel my shoulders are tight and my legs want to run."

As the therapist was interacting with Molly, she noted that not only had her voice changed but the pressure in her speech was less. She would glance at the therapist, and she appeared almost curious about the questions. At other points in Molly's therapy, in the group, she had used an intervention which she considered at this juncture, always with Molly's explicit agreement. The therapist would offer some words that Molly might say about her experience, and some simple physical actions to make with the words that amplified their meaning. And with the other patients' agreement, Molly would express herself in the other patients' direction, to facilitate a sense of relational interaction, and evidence the other holds no malice. Such interventions had helped Molly decrease her habitual response, by increasing her felt sense of her body, supporting her increased awareness of what emotion she was feeling in the moment.

The therapist asked Molly if she was willing to "try something to help her continue moving out of that bound up place," to which Molly said "yes." She invited Molly to sit out further in her chair and look at the patient who had addressed her earlier in the session.

"How about you say to her: "You don't know what it is to not have help!" As loudly as you can."

Molly began hesitantly, barely able to keep eye contact, but gamely trying to do so. With encouragement Molly's voice became louder with each attempt. Soon she was saying the phrase very loudly, assertively and with direct eye contact. She then became quiet and tears began to flow readily down her face, the frown gone and her eyes much more expressive. All the members of her group were looking at her with encouragement, even the person at whom she was yelling. The tension in Molly's body lessened visibly.

"Can you feel any more sensations now?" Molly replied with a wry chuckle as she motioned across the room, "My eyes feel less tight, and she doesn't look like she did a while ago. I don't know why but my head doesn't have that bound up feeling, my legs feel really tingly, and my chest – it feels like there is something moving in it. And I feel a deep pain there." She sat back with her eyes softening more, appearing to consider something carefully inside herself. "Well, the feeling that there isn't any help has lessened, and I don't feel so alone, but that tightness in my forehead is a lot less and some sort of feeling in my chest is really strong now. I think it feels like my heart has a place in my chest now? Which is good I know, but wow, does it hurt, too."

During the subsequent interaction with another group member, Molly's habitual prior is voiced that others cannot be counted on, and must be critical of her. She became very agitated and after dismissing the other as "*an idiot*," she said she was "*pissed*." However, her tears and physical agitation were physical signals that she was not likely to be referencing the current interoceptive processes that were instigating feelings. When asked, why she was feeling what she claimed she felt, Molly further elaborated on the ruminative material ("*pissed*"), as the precision afforded to such beliefs severely limited any possible awareness of other options. As the therapist inquired into her bodily state, she helped Molly place her attention on her reported "*tightness*" and decrease her habitual ruminations. She understood that Molly couldn't, at that moment, question the content of what the other (compassionate) patient had said to her. Instead, the therapist began drawing attention to the sequelae of her (overdetermined) sympathetic nervous system reaction of the moment, noted as tension by Molly, asking "*What do you feel in your body right now?*"

The therapist began with Molly's word choice ("*tightness*" and "*bound up*") but increasingly expanded the field of options ("*vulnerable*" and "*danger*| "). As she again helped Molly to direct her attention to different aspects of the experience, this supported Molly in gaining control of her attentional resources. This is crucial in making her interoceptive sensations more precise (e.g., "*Can you say where the tightness is most*?"). While this initially increased the strength of the affective feeling (as evidenced by Molly ducking her head and insisting that she has to withdraw into herself, "*where I know I'm safe*"), disconfirming evidence can gain precision by this path and new alternative priors can be considered. At this point the real relationship between the therapist and Molly is an anchoring element. The manner in which the therapist readily accepts Molly's description at first but continues to question, as well as the inflection of her voice ("*You would be vulnerable if they saw your feelings, eh*?"), reflect that she and Molly have been through such moments before. Molly's decreasing tension implies a change in the experienced emotional valence to a more positive level (even though she is uncertain of the outcome), implying increasing confidence that they have made it through together before to where the therapist is leading, without encountering the dangers she is habitually expecting at that moment.

As Molly continued to engage, her voice became less reactive, and she even expressed curiosity. As the therapist observed Molly's increasing self-reflective stance, she decided it would be clinically helpful to invite Molly to express her experience to others in ways that lessened her sense of aloneness, speaking directly to the other patients, who verbalized their willingness.

The intervention with the other patient in the group encouraged Molly to address her prior beliefs that she was unsafe in interaction with others. She was helped to shift from the position of a "stubborn scientist" to that of a Bayesian observer. The physical changes in her body, as well as the increase in her verbal deliberations about the possible implications of such experience in the moment, is evidence of changes in the precisions of sensations throughout the hierarchy which activated her body and had previously prevented the evaluation of new evidence. Ultimately, at that point she could recognize others' presence with her and allow such presence to be helpful. While there would have to be many more episodes of such experimental trials, as precisions develop a life of their own, Molly was able to acknowledge aloud a sense of vulnerability and "*deep pain*." Her feelings could be witnessed by others, which she recognized as safe and necessary, indicating a more emotionally open and deliberative stance. Molly was then able to make a profound statement about the embodied experience of her own emotion – as if her heart had gained space in her chest, which was experienced as both somewhat painful yet also soothing.

Molly's response to the therapist's recommendation to engage with the other patients also highlights various findings in

neuroscience research. Molly's voice became more assertive with each attempt and she was able to direct her eye contact purposefully. Eye contact has been found to support attention to subjective experience and increase the accuracy of emotional report about interoceptive experience (Baltazar et al., 2014). Molly used more expressive language ("*I don't feel so alone*") to describe her feeling state (i.e., increasing the precision of interoceptive sensation and prediction errors, to counter over-precise priors). Allowing others to witness her tears is indicative of lessening fear in Molly, reflecting a changing emotional valence for her, and likely falling free energy. Making eye contact, breathing deeper and crying openly is proof that she was experiencing a change in emotional valence, suggestive of an increasing positive valence, as she became more hopeful. Such behavior and the resulting interactions between Molly and her therapist became possible, as free energy decreased, lessening the need for the highly defensive behavior exhibited before. Changes in her affective experience as Molly sat up straighter in her seat reflect Ceunen et al. (2014) finding that an upright posture is associated with more positive affect than slouching, possibly because an upright stance expresses pride or power. During the intervention in the group, there was an obvious shift in Molly's ANS reactivity, as evidenced by her chuckle and less bodily tension (Ceunen et al., 2014). Payne et al. (2015) assert that diverting the patient's attention temporarily to a physical experience that gives them a sense of safety can mitigate extremes in autonomic reactivity in the body (Payne et al., 2015). Then – a little at a time – the patient can turn their attention again to the disturbance and slowly regain ANS balance. For Molly this intervention ostensibly created the physiologic equivalent of emotional "space" from which she could more deliberately address her experience in the present moment. She can consider alternatives to her prior beliefs of danger in the world, even ultimately allowing the open expression of vulnerability, with tears.

CONCLUSION

Our patients come to therapy when habitual responses, which are embedded within their physiology, fail to produce expected or desired outcomes. Predictive processing theories of the brain as an inference machine cast valuable light on how such dysfunctional patterns of responding can come about in infancy and be highly resistant to change. In order to change a prior it is necessary to act on the interoceptive system that created that prior in the first place.

Increasing attention to interoceptive sensation changes the balance of precision between the current interoceptive sensation and the "stubborn prior." This change in precision can update a resistant prior and in doing so increase the patient's ability to "mentalize interoception," allowing alternative hypotheses to be generated about subjective experience. Intervening to influence precision similarly supports the patient's efforts to bring emotion into awareness, which increases opportunities for their verbal expression – an important outcome of any therapeutic encounter.

We propose that the crucial point of access, within the therapeutic relationship is for the patient to focus attention onto their current internal bodily sensations (their interoception). Attention to the body, and the feelings that accompany this, sets in train a series of responses that may permit updating of default/habitual beliefs and the expectations that cause the patient distress in their current relationship to themselves, others and the world. We describe how this can re-calibrate the patient's interoceptive responses, increase emotional awareness, strengthen evaluative thought patterns and allow the patient the flexibility to discern what is real and present in any given moment.

AUTHOR CONTRIBUTIONS

PD initiated the manuscript. Both authors contributed equally to the composition of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers (KF and MM) and Jim Hopkins, for their help in delineating the above ideas; and Cynthia Duquette for editorial support.

REFERENCES

Adam, G. (2010). *Visceral Perception: Understanding Internal Cognition*. New York, NY: Plenum Press.

Adolfi, F., Couto, B., Richter, F., Decety, J., Lopez, J., Sigman, M., et al. (2017). Convergence of interoception, emotion, and social cognition: a twofold fMRI meta-analysis and lesion approach. *Cortex* 88, 124–142. doi: 10.1016/j.cortex.2016.12.019

Allen, J. G., Fonagy, P., and Bateman, A. W. (2008). *Mentalizing in Clinical Practice*. Washington, D.C: American Psychiatric Publishing.

Allen, M., Levy, A., Parr, T., and Friston, K. (2019). In the body's eye: the computational anatomy of interoceptive inference. *Biorxiv* 10, doi: 10.1101/603928

Baltazar, M., Hazem, N., Vilarem, E., Beaucousin, V., Picq, J. L., and Conty, L. (2014). Eye contact elicits bodily self-awareness in human adults. *Cognition* 133, 120–127. doi: 10.1016/j.cognition.2014.06.009

Bar-Levav, R. (1988). *Thinking in the Shadow of Feelings: A New Understanding of the Hidden Forces That Shape Individuals and Societies*. New York, NY: Simon and Schuster.

Barrett, L. F. (2011). Constructing emotion. *Psychol. Top.* 20, 359–380.

Barrett, L. F., Quigley, K., Bliss-Moreau, E., and Aronson, K. (2004). Interoceptive sensitivity and self-reports of emotional experience. *J. Personal. Soc. Psychol.* 87, 684–697. doi: 10.1037/0022-3514.87.5.684

Barrett, L. F., Quigley, K. S., and Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philos. Trans. R. Soc. B* 371:20160011. doi: 10.1098/rstb.2016.0011

Barrett, L. F., and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* 16, 419–429. doi: 10.1038/nrn3950

Bateman, A. W., and Fonagy, P. (2012). *Handbook of Mentalizing in Mental Health Practice*. Washington, D.C: American Psychiatric Publishing.

Berntson, G. G., Norman, G. J., Bechara, A., Bruss, J., Tranel, D., and Cacioppo, J. T. (2011). The insula and evaluative processes. *Psychol. Sci.* 22, 80–86. doi: 10.1177/0956797610391097

Besharati, S., Forkel, S. J., Kopelman, M., Solms, M., Jenkinson, P. M., and Fotopoulou, A. (2015). Mentalising the body: spatial and social cognition in anosognosia for hemiplegia. *Brain* 139(Pt 3), 971–985. doi: 10.1093/brain/awv390

Bowlby, J. (1988). *A Secure Base: Parent Child Attachment and Healthy Human Development*. London: Routledge.

Bruineberg, J., Kiverstein, J., and Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese* 195, 2417–2444. doi: 10.1007/s11229-016-1239-1231

Cameron, O. G. (2001). Interoception: the inside story–a model for psychosomatic processes. *Psychosom. Med.* 63, 697–710. doi: 10.1097/00006842-200109000-00001

Cannon, W. B. (1932). *The Wisdom of the Body*. New York, NY: W.W. Norton & Company.

Ceunen, E., Vlaeyen, J. W., and Van Diest, I. (2016). On the Origin of Interoception. *Front. Psychol.* 23:743. doi: 10.3389/fpsyg.2016.00743

Ceunen, E., Zaman, J., Vlaeyen, J. W., Dankaerts, W., and Van Diest, I. (2014). Effect of seated trunk posture on eye blink startle and subjective experience: comparing flexion, neutral upright posture, and extension of spine. *PLoS One* 9:e88482. doi: 10.1371/journal.pone.0088482

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Cozolino, L. (2002). *The Neuroscience of Psychotherapy: Building and Rebuilding the Human Brain*. New York, NY: W.W. Norton & Company.

Craig, A. D. (2002).). How do you feel? Interoception: the sense of the physiological condition of the body. *Nat. Rev. Neurosci.* 3, 655–666. doi: 10.1038/nrn894

Craig, A. D. (2008). "Interoception and emotion: a neuroanatomical perspective," *Handbook of Emotions,* 3 Edn, eds Lewis, M, Haviland-Jones, J, Feldman Barrett, L (New York, NY: Guilford Press), 272-290.

Craig, A. D. (2009a). Emotional moments across time: a possible neural basis for time perception in the anterior insula. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 1933–1942. doi: 10.1098/rstb.2009.0008

Craig, A. D. (2009b). How do you feel — now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10, 59–70. doi: 10.1038/nrn2555

Craig, A. D. (2010). The sentient self. *Brain Struct. Funct.* 6, 563–577. doi: 10.1007/s00429-010-0248-y

Craig, A. D. (2011). Significance of the insula for the evolution of human awareness of feelings from the body. *Ann. N. Y. Acad. Sci.* 1225, 72–82. doi: 10.1111/j.1749-6632.2011.05990.x

Craig, A. D. (2015). *How Do You Feel? An Interoceptive Moment with Your Neurobiological Self*. Oxford: Princeton University Press.

Critchley, H. D., and Garfinkel, S. N. (2017). Interoception and emotion. *Cur.r Opin. Psychol.* 17, 7–14. doi: 10.1016/j.copsyc.2017.04.020

Critchley, H. D., and Harrison, N. A. (2013). Visceral Influences on Brain and Behavior. *Neuron* 77, 624–638. doi: 10.1016/j.neuron.2013.02.008

Critchley, H. D., and Nagai, Y. (2012). How emotions are shaped by bodily states. *Emot. Rev.* 4, 163–168. doi: 10.1177/1754073911430132

Critchley, H. D., and Seth, A. (2012). Will studies of macaque insula reveal the neural mechanisms of self-awareness? *Neuron* 74, 423–426. doi: 10.1016/j.neuron.2012.04.012

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., and Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nat. Neurosci.* 7, 189–195. doi: 10.1038/nn1176

Damasio, A., Damasio, H., and Tranel, D. (2013). Persistence of feelings and sentience after bilateral damage of the insula. *Cereb. Cortex* 23, 833–846. doi: 10.1093/cercor/bhs077

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. New York, NY: Picador.

Damasio, A. R. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. New York, NY: Pantheon Books.

Daw, N. (2015). Of goals and habits. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13749–13750. doi: 10.1073/pnas.1518488112

Dezfouli, A., and Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *Eur. J. Neurosci.* 35, 1036–1051. doi: 10.1111/j.1460-9568.2012.08050.x

Dolan, R. J., and Dayan, P. (2013). Goals and habits in the brain. *Neuron* 80, 312–325. doi: 10.1016/j.neuron.2013.09.007

Duquette, P. (2010). Reality matters: attachment, the real relationship, and change in psychotherapy. *Am. J. Psychother.* 64, 127–151. doi: 10.1176/appi.psychotherapy.2010.64.2.127

Duquette, P. (2017). Increasing our insular world view: interoception and psychopathology for psychotherapists. *Front. Neurosci.* 21:135. doi: 10.3389/fnins.2017.00135

Edwards, M. J., Adams, R. A., Brown, H., Parees, I., and Friston, K. J. (2012). A Bayesian account of 'hysteria'. *Brain* 135, 3495–3512. doi: 10.1093/brain/aws129

Eshkevari, E., Rieger, E., Musiat, P., and Treasure, J. (2014). An investigation of interoceptive sensitivity in eating disorders using a heartbeat detection task and a self-report measure. *Eur. Eat. Disord. Rev.* 22, 383–388. doi: 10.1002/erv.2305

Farb, N., Daubenmier, J., Price, C. J., Gard, T., Kerr, C., Dunn, B. D., et al. (2015). Interoception, contemplative practice, and health. *Front. Psychol.* 9:763. doi: 10.3389/fpsyg.2015.00763

Feldman, H., and Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* 2:215. doi: 10.3389/fnhum.2010.00215

Fletcher, P. C., and Frith, C. D. (2009). Perceiving Is Believing: a Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia. *Nat. Rev. Neurosci.* 10, 48–58. doi: 10.1038/nrn2536

Fogel, A. (2009). *The Psychophysiology of Self-Awareness: Rediscovering the Lost Art of Body Sense*. New York, NY: W.W. Norton.

Fonagy, P., and Target, M. (2006). The mentalization-focused approach to self pathology. *J. Pers. Disord.* 20, 544–576. doi: 10.1521/pedi.2006.20.6.544

Fotopoulou, A. (2015). The virtual bodily self: mentalization of the body as revealed in anosognosia for hemiplegia. *Conscious. Cogn.* 33, 500–510. doi: 10.1016/j.concog.2014.09.018

Fotopoulou, A., and Tsakiris, M. (2017). Mentalizing homeostasis: the social origins of interoceptive inference. *Neuropsychoanalysis* 19, 3–28. doi: 10.1080/15294145.2017.1294031

Frijda, N. H. (1986). *The Emotions*. Cambridge, MA: Cambridge University Press.

Frijda, N. H. (2007). *The Laws of Emotion*. Mahwah, NJ: Lawrence Erlbaum Associate.

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475. doi: 10.1098/rsif.2013.0475

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., and Pezzulo, G. (2016). Active inference and learning. *Neuro. Biobehav. Rev.* 68, 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Friston, K. J., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol.* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Frith, C. (2007). *Making up the Mind: How the Brain Creates Our Mental World*. Hoboken, NJ: Wiley & Sons, 2007.

Gelso, C. J. (2009). The real relationship in a postmodern world: theoretical and empirical explorations. *Psychother. Res.* 19, 253–264. doi: 10.1080/10503300802389242

Gu, X., and FitzGerald, T. (2014). Interoceptive Inference: homeostasis and Decision-making. *Trends Cogn. Sci.* 18, 269–270. doi: 10.1016/j.tics.2014.02.001

Gu, X., Gao, Z., Wang, X., Liu, X., Knight, R. T., Hof, P. R., et al. (2012). Anterior insular cortex is necessary for empathetic pain perception. *Brain* 135, 2726–2735. doi: 10.1093/brain/aws199

Gu, X., Hof, P. R., Friston, K., and Fan, J. (2013a). Anterior insular cortex and emotional awareness. *J. Comp. Neurol.* 521, 3371–3388. doi: 10.1002/cne.23368

Gu, X., Liu, X., Van Dam, N. T., Hof, P. R., and Fan, J. (2013b). Cognition-emotion integration in the anterior insular cortex. *Cereb. Cortex* 23, 20–27. doi: 10.1093/cercor/bhr367

Harrison, N. A., Gray, M. A., Gianaros, P. J., and Critchley, H. D. (2010). The embodiment of emotional feelings in the brain. *J. Neurosci.* 38, 12878–12884. doi: 10.1523/JNEUROSCI.1725-10.2010

Harshaw, C. (2015). Interoceptive dysfunction: toward an integrated framework for understanding somatic and affective disturbance in depression. *Psychol. Bull.* 141, 311–363. doi: 10.1037/a0038101

Herbert, B. M., and Pollatos, O. (2012). The body in the mind: on the relationship between interoception and embodiment. *Top. Cogn. Sci.* 4, 692–704. doi: 10.1111/j.1756-8765.2012.01189

Hoemann, K., Gendron, M., and Barett, L. (2017). Mixed emotions in the predictive brain. *Curr. Opin. Behav. Sci.* 15, 51–57. doi: 10.1016/j.cobeha.2017.05.013

Hyett, M., Parker, G. B., Guo, C., Zalesky, A., Nguyen, V. T., Yuen, T., et al. (2015). Scene unseen: disrupted neuronal adaptation in melancholia during emotional film viewing. *Neuroimage Clinical.* 9, 660–667. doi: 10.1016/j.nicl.2015.10.011

James, W. (1890). *The Principles of Psychology*. New York, NY: Holt.

Joffily, M., and Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Comput. Biol.* 9:e1003094. doi: 10.1371/journal.pcbi.1003094

Jones, C. L., Minati, L., Nagai, Y., Medford, N., Harrison, N. A., Gray, M., et al. (2015). Neuroanatomical substrates for the volitional regulation of heart rate. *Front. Psychol.* 6:300. doi: 10.3389/fpsyg.2015.00300

Khalsa, S., and Lapidus, R. (2016). Can interoception improve the pragmatic search for biomarkers in psychiatry? *Front. Psychiatry* 7:121. doi: 10.3389/fpsyt.2016.00121

Kube, T., Rief, W., and Glombiewski, J. A. (2017). On the maintenance of expectations in major depression – Investigating a neglected phenomenon. *Front. Psychol.* 8:9. doi: 10.3389/fpsyg.2017.00009

Kube, T., Schwarting, R., Rozenkrantz, L., Glombiewski, J. A., and Rief, W. (2019). Distorted cognitive processes in major depression – A predictive processing perspective. *Biol. Psychiatry* (in press). doi: 10.1016/j.biopsych.2019.07.017

LeDoux, J. (2002). *Synaptic Self: How Our Brains Become Who We Are*. New York, NY: Viking.

Lyons-Ruth, K., Brunshweiler-Stern, N., Harrison, A. M., Morgan, A. M., Nahum, J. P., Sander, L., et al. (1998). Implicit relational knowing: Its role in development and psychoanalytic treatment. *Infant Ment. Health J.* 19, 282–289. doi: 10.1002/(sici)1097-0355(199823)19:3<282::aid-imhj3>3.3.co;2-f

Medford, N., and Critchley, H. D. (2010). Conjoint activity of anterior insular and anterior cingulate cortex: awareness and response. *Brain Struct. Funct.* 214, 535–549. doi: 10.1007/s00429-010-0265

Miller, M., and Clark, A. (2018). Happily entangled: prediction, emotion, and the embodied mind. *Synthese* 195, 2559–2575. doi: 10.1007/s11229-017-1399-1397

Morgan, A. C., Brunshweiler-Stern, N., Harrison, A. M., Lyons-Ruth, K., Nahum, J. P., Sander, L., et al. (1998). Moving along to things left undone. *Infant Ment. Health J.* 19, 324–332. doi: 10.1002/(sici)1097-0355(199823)19:3<324::aid-imhj9>3.0.co;2-l

Northoff, G. (2016). *Neuro-Philosophy and the Healthy Mind: Learning from the Unwell Brain*, 1 Edn. New York, NY: Norton.

Ondobaka, S., Kilner, J., and Friston, K. J. (2017). The role of interoceptive inference in theory of mind. *Brain Cogn.* 112, 64–68. doi: 10.1016/j.bandc.2015.08.002

Owens, A. P., Allen, M., Ondobaka, S., and Friston, K. (2018). Interoceptive inference: from computational neuroscience to clinic. *Neurosci. Biobehav.* 90, 174–183. doi: 10.1016/j.neubiorev.2018.04.017

Owens, A. P., David, A. S., Low, D. A., Mathias, C. J., and Sierra-Siegert, M. (2015). Abnormal cardiovascular sympathetic and parasympathetic responses to physical and emotional stimuli in depersonalization disorder. *Front. Neurosci.* 9:89. doi: 10.3389/fnins.2015.00089

Panskepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions (Series in Affective Science)*, 1st Edn. New York, NY: Oxford University Press.

Paulus, M. P., and Stein, M. B. (2006). An insular view of anxiety. *Biological. Psychiatry* 60, 383–387. doi: 10.1016/j.biopsych.2006.03.042

Paulus, M. P., and Stein, M. B. (2010). Interoception in anxiety and depression. *Brain Struct. Funct.* 214, 451–462. doi: 10.1007/s00429-010-0258-259

Payne, P., Levine, P. A., and Crane-Godreau, M. A. (2015). Somatic experiencing: using interoception and proprioception as core elements of trauma therapy. *Front. Psychol.* 6:93. doi: 10.3389/fpsyg.2015.00093

Petzschner, H., Weber, L., Gard, T., and Stephan, K. E. (2017). Computational psychosomatics and computational psychiatry: toward a joint framework for differential diagnosis. *Biol. Psychiatry* 82, 421–430. doi: 10.1016/j.biopsych.2017.05.012

Pezzulo, G. (2014). Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cogn. Affect. Behav. Neurosci.* 14, 902–911. doi: 10.3758/s13415-013-0227

Quadt, L., Critchley, H. D., and Garfinkel, S. (2018). The neurobiology of interoception in health and disease. *Ann. N.Y. Acad. Sci.* 1428, 112–126.

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubo, A., Schwarting, R., and Torwart, A. (2015). Expectancies are core feature of mental disorders. *Curr. Opin. Psychiatry* 28, 378–385. doi: 10.1097/yco.0000000000000184

Rolls, E. T. (1999). *The Brain and Emotion*. Oxford: Oxford University Press.

Russell, J. A., and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *J. Pers Soc. Psych.* 76, 805–819. doi: 10.1037/0022-3514.76.5.805

Schachter, S., and Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychol. Rev.* 69, 379–399. doi: 10.1037/h0046234

Schore, A. N. (2003). *Affect Regulation and the Repair of the Self*. New York, NY: W.W. Norton & Company.

Schore, A. N. (2012). *The Science of the Art of Psychotherapy. Norton Series on Interpersonal Neurobiology*. New York, N.Y: W.W. Norton.

Sedeno, L., Couto, B., Melloni, M., Canales-Johnson, A., Yoris, A., and Baez, S. (2014). How do you feel when you can't feel your body? Interoception, functional connectivity and emotional processing depersonalization-derealization disorder. *PLoS One* 9:e98769. doi: 10.1371/journal.pone.0098769

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* 17, 565–573. doi: 10.1016/j.tics.2013.09.007

Seth, A. K., and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philos. Trans. R. Soc. B* 371:20160007. doi: 10.1098/rstb.2016.0007

Seth, A. K., Suzuki, K., and Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Front. Psychol.* 2:395 doi: 10.3389/fpsyg.2011.00395

Siegel, D. J. (1999). *The Developing Mind: Toward a Neurobiology of Interpersonal Experience*. New York, NY: Guilford Press.

Siegel, D. J. (2010). *The Mindful Therapist*. New York, NY: W.W. Norton & Company.

Singer, T., Critchley, H. D., and Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends Cogn. Sci.* 13, 334–340. doi: 10.1016/j.tics.2009.05.001

Solms, M. (2019). The hard problem of consciousness and the free energy principle. *Front. Psychol.* 9:2714. doi: 10.3389/fpsyg.2018.02714

Stephan, K. E., Manjaly, Z., Mathys, C., Weber, L., Paliwal, S., Gard, T., et al. (2016). Alostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Front. Hum. Neurosci.* 10:550. doi: 10.3389/fnhuman.2016.00550

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiol. Behav.* 106, 5–15. doi: 10.1016/j.physbeh.2011.06.004

Sterling, P. (2014). Homeostasis vs allostasis: implications for brain function and mental disorders. *JAMA Psychiatry* 71, 1192–1193.

Stern, D. (1985). *The Interpersonal World of the Infant: A View From Psychoanalysis and Developmental Psychology*. New York, NY: Basic Books.

Stern, D., Bruschweiler-Stern, N., Harrison, A. M., Lyons-Ruth, K., Morgan, A. C., Nahum, J. P., et al. (1998). The process of therapeutic change involving implicit knowledge: some implications of developmental observations for adult psychotherapy. *Infant Ment. Health J.* 19, 300–308. doi: 10.1002/(sici)1097-0355(199823)19:3<300::aid-imhj5>3.0.co;2-p

Strigo, I. A., and Craig, A. D. (2016). Interoception, homeostatic emotions and sympathovagal balance.". *Philos. Trans. R. Soc. Lond. B Biol. Sci.*371:20160010. doi: 10.1098/rstb.2016.0010

Vlemincx, E., van Diest, I., and van den Bergh, O. (2015). Emotion, sighing, and respiratory variability. *Psychophysiology* 52, 657–666. doi: 10.1111/psyp.12396

Wallin, D. J. (2007). *Attachment in Psychotherapy*. New York, NY: The Guilford Press.

Wiens, S. (2005). Interoception in emotional experience. *Curr. Opin. Neurol.* 18, 442–447. doi: 10.1097/01.wco.0000168079.92106.99

Yon, D., de Lange, F. P., and Press, C. (2019). The predictive brain as a stubborn scientist. *Trends Cogn. Sci.* 23, 6–8. doi: 10.1016/j.tics.2018.10.003

Zaki, J., Davis, J., and Ochsner, K. (2012). Overlapping activity in anterior insula during interoception and emotional experience. *NeuroImage* 62, 493–499. doi: 10.1016/j.neuroimage.2012.05.012

Zautra, A. J., Fasman, R., Davis, M. C., and Craig, A. D. (2010). The effects of slow breathing on affective responses to pain stimuli: an experimental study. *Pain* 149, 12–18. doi: 10.1016/j.pain.2009.10.001

Conflict of Interest: 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 Psychology

HYPOTHESIS AND THEORY
published: 04 October 2019
doi: 10.3389/fpsyg.2019.02267

The Predictive Processing Model of EMDR

*D. Eric Chamberlin**

Chamberlin Applied Neuroscience, Glastonbury, CT, United States

Eye Movement Desensitization and Reprocessing Therapy (EMDR) is an effective treatment for Post-traumatic Stress Disorder (PTSD). The Adaptive Information Processing Model (AIP) guides the development and practice of EMDR. The AIP postulates inadequately processed memory as the foundation of PTSD pathology. Predictive Processing postulates that the primary function of the brain is prediction that serves to anticipate the next moment of experience in order to resist the dissipative force of entropy thus facilitating continued survival. Memory is the primary substrate of prediction, and is optimized by an ongoing process of precision weighted prediction error minimization that refines prediction by updating the memories on which it is based. The Predictive Processing model of EMDR postulates that EMDR facilitates the predictive processing of traumatic memory by overcoming the bias against exploration and evidence accumulation. The EMDR protocol brings the traumatic memory into an active state of re-experiencing. Defensive responding and/or low sensory precision preclude evidence accumulation to test the predictions of the traumatic memory in the present. Sets of therapist guided eye movements repeatedly challenge the bias against evidence accumulation and compel sensory sampling of the benign present. Eye movements reset the theta rhythm organizing the flow of information through the brain, facilitating the deployment of both overt and covert attention, and the mnemonic search for associations. Sampling of sensation does not support the predictions of the traumatic memory resulting in prediction error that the brain then attempts to minimize. The net result is a restoration of the integrity of the rhythmic deployment of attention, a recalibration of sensory precision, and the updating (reconsolidation) of the traumatic memory. Thus one prediction of the model is a decrease in Attention Bias Variability, a core dysfunction in PTSD, following successful treatment with EMDR.

Keywords: psychological trauma (PTSD), Free Energy Principle, predictive processing, EMDR, memory reconsolidation, physiological mechanism

OPEN ACCESS

Edited by:
Christoph Mathys, International School for Advanced Studies (SISSA), Italy

Reviewed by:
Karl Friston, University College London, United Kingdom
Gerd Thomas Waldhauser, Ruhr University Bochum, Germany

***Correspondence:**
D. Eric Chamberlin
Eric@ChamberlinNeuroscience.com

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *10 July 2019*
Accepted: *23 September 2019*
Published: *04 October 2019*

Citation:
Chamberlin DE (2019) The Predictive Processing Model of EMDR. Front. Psychol. 10:2267. doi: 10.3389/fpsyg.2019.02267

INTRODUCTION

The Adaptive Information Processing Model (AIP) guides the development and practice of Eye Movement Desensitization and Reprocessing Therapy (EMDR) used in the treatment of Post-traumatic Stress Disorder. The AIP hypothesizes that "dysfunctionally stored memory" serves as the foundation of post-traumatic psychopathology (Shapiro, 2001). Furthermore "there is a system inherent in all of us that is physiologically geared to process information to a state of mental health. . . by means of this system, negative emotions are relieved, and learning takes place, is

appropriately integrated, and is available for future use" (Shapiro, 2018). EMDR is posited to exert its therapeutic effects through targeted information processing of "dysfunctionally stored memory" (Solomon and Shapiro, 2008; Shapiro and Laliotis, 2011; Shapiro, 2018).

The clinical effectiveness of EMDR has been well-established (Rodenburg et al., 2009; Bisson et al., 2013; Jeffries and Davis, 2013; Watts et al., 2013). However the proposed neurobiological mechanisms of EMDR have yet to offer a model capable of catalyzing robust targeted biological research (Bergmann, 2010). To wit, in a recent review Landin-Romero et al. (2018) concluded, "the current understanding of the mechanisms of action underlying EMDR is similar to the parable of the Blind Men and the Elephant in that there is no agreed definition of what the candidate mechanisms are (i.e., eye movements, bilateral stimulation, dual attention, etc.) and how these mechanisms can be measured or demonstrated." The goal of this paper is to try to remedy this situation through application of Predictive Processing to EMDR.

THE FREE ENERGY PRINCIPLE – FOUNDATION OF PREDICTIVE PROCESSING

Predictive Processing is a corollary of the Free Energy Principle as developed by Friston et al. (2006) and Friston (2009). The Free Energy Principle has its roots in statistical physics as an information isomorph of the second law of thermodynamics (Clark, 2013c). Living requires energy and information to maintain organization and resist the dispersive forces of entropy. For humans this requires statistical processing that minimizes uncertainty about the world, despite not having direct access to the world. As postulated by the Free Energy Principle, this is accomplished through a generative model that is constantly updated to reflect current conditions (Friston et al., 2006). The existence and updating of a generative model is where the Predictive Processing story begins.

PREDICTIVE PROCESSING

The activity of "boot strapping" increasingly complex models of the world based on probabilistic inference is known as Predictive Processing (Clark, 2013c). From the perspective of Predictive Processing the main function of the brain is to predict its own immediate experience, i.e., the patterns of firing neurons that will occur next. To achieve this goal there is a relentless focus on reducing the errors of its predictions so as to "get it right" in the future. This process is known as Prediction Error Minimization and utilizes sensation as feedback on the accuracy of its predictions. The excitement surrounding Predictive Processing in contemporary neuroscience stems from the promise of being able to explain a wide range of cognitive activities including perception, attention, learning, and action with a single conceptually simple mechanism grounded in physiologically plausible computation (Friston, 2009; Hohwy, 2013). Recently this paradigm has been applied to Psychotherapy (Holmes and Nolte, 2019).

THE PREDICTIVE PROCESSING MODEL OF EMDR

The Predictive Processing Model of EMDR focuses on the role of memory as the principle substrate for predictions that guide behavior (Bar, 2009; Buckner, 2010). To minimize uncertainty, resist entropy, and ensure survival the brain is constantly making predictions, and then using sensation as feedback to test its predictions (Clark, 2013c). When there is a mismatch between what is predicted and what is currently sensed, the brain registers a "prediction error" (Hohwy, 2013). In response the brain may update the memory through the process of Memory Reconsolidation (Pedreira et al., 2004; Dudai, 2006, 2009). The goal of updating the memory is to minimize the (long-term average of) prediction error thus reducing uncertainty and resulting in more successful behavior in the future. The dysfunctionally stored memories postulated by the AIP make for poor predictions and result in the suboptimal behavior characteristic of PTSD. Thus The Predictive Processing Model of EMDR attempts to explain the biological basis of "the system inherent in all of us that is physiologically geared to process information to a state of mental health" postulated by the AIP and activated by EMDR.

Eye Movement Desensitization and Reprocessing Therapy is an ideal lens through which to view this model as it is "a comprehensive psychotherapy compatible with all theoretical orientations," and has well-delineated clinical interventions (Shapiro and Laliotis, 2011; Shapiro, 2018). In addition the inclusion of the therapeutic element of eye movements affords the opportunity to appreciate the powerful role that eye movements play in network and mnemonic function (Johansson and Johansson, 2014; Vernet et al., 2014).

PERCEPTION AS INFERENCE

Predictive Processing has its roots in the work of German physician and physicist Herman von Helmholtz (Friston et al., 2006). Helmholtz recognized that incoming sensory data are ambiguous (Helmholtz, 1867). For a given sensation there are multiple potential causes in the world. For example, an orange scent could be caused by orange soda, air freshener, or an actual orange. And contrary to common sense, we do not have direct access to the world. Consider vision. Light does not enter the brain. The inside of the skull is dark. Instead the retina converts photons of light into the firing of neurons. In fact every sensory receptor, from vision, to touch, to smell, has the same type of output. This is true for the interoceptive senses such as proprioception, hunger, and thirst as well. In our experience of the world, all the brain has to work with are patterns of firing neurons. As Immanuel Kant suggested, all we can know is the "phenomenon," that is the effect of the world upon us, i.e., patterns of firing neurons. We can never know

"the thing in itself" that is, the actual causes in the world of the effects we experience (Kant, 1781). With this observation Kant anticipated the Markov Blanket, a concept central to Predictive Processing. The Markov blanket is essentially the boundary between a system, and everything else that is not that system, expressed in mathematical terms (Yufik and Friston, 2016). Given the ambiguity of sensory data and the impossibility of knowing "the thing in itself" Helmholtz concluded that perception is an act of unconscious inference. We cannot know directly what lies on the other side of the Markov blanket (i.e., sensory boundary) that is constituted by our sensory epithelia. When we perceive, the brain is making a guess about the state of the world. This process is automatic, rapid, and unconscious (Tenenbaum et al., 2011). As a result we are unaware that a sophisticated process has occurred. We are only aware of the product, what the brain has calculated is the most likely cause, the best guess. However we do not experience this as a probability or a guess, but rather as a fact (Dehaene and Changeux, 2011; Hohwy, 2013) "I see an orange."

Helmholtz' hypothesis of perception as inference has significant implications for brain function. If perception is an act of inference, the brain must have information that is used as the basis for inference. That is, it must have a model of the world, *a priori*, before it encounters the world. Dreaming during Rapid Eye Movement (REM) sleep illustrates the ability of the brain to generate perceptual hypotheses in the absence of any sensory data, an *a priori* model (Hobson et al., 2014). In the parlance of predictive processing this is called a prior probability or "prior" based on Bayes Theorem (Geisler and Diehl, 2003). Prior probability is the likelihood of a proposition before considering empirical data from the senses. But where does such a prior probability or model come from?

HARDWIRED MODELS

The models present at birth appear to be hardwired (Ullman et al., 2012). As suggested by Kant, in order for humans to be able to make sense of the world we assume that experience unfolds in extended space, over time, with causes and effects (Kant, 1781). In other words the hardwired model we begin life with contains notions of space, time and causality. Friston has suggested that hardwired models are a function of the type of organism, including its particular sensory receptors and expected environment. Biological systems have a model implicit in their structure, and sample the world so as to fulfill their expectations (Friston et al., 2006). Fish "expect" to be surrounded by water from which they extract oxygen. Humans "expect" to be surrounded by air. Such "proto-concepts" form the scaffolding upon which patterns of firing neurons resulting from experience give rise to more sophisticated models of the world (Ullman, 2019).

EVOLUTION OF MODELS

Following birth humans "boot strap" increasingly complex models of the world based on experience. Prior probabilities present at birth are modified by experience into posterior probabilities. For example, starting with no knowledge of language, infants learn language. Research is beginning to deconstruct this process through the lens of predictive processing. One of the first tasks of an infant learning language is to parse a stream of syllables into discrete words. Based on patterns of firing neurons from the cochlea, the infant identifies some combinations of sounds as occurring more frequently together than others. Given 2 min of exposure, 8 month old infants can separate the syllables Pre-tty-ba-by into the separate words of pretty, and baby (Saffran et al., 1996). The syllables pre-tty occur more frequently together in natural speech than the syllables found in the middle of the stream, tty-ba. Similarly the syllables ba-by occur more frequently together than tty-ba. The infant's best guess based on statistical computations about the patterns of firing neurons that it experiences is that "pretty" is a discrete word, and that "baby" is a discrete word. A prior probability that speech sounds that occur in particular patterns have significance, becomes a posterior probability that "pretty" and "baby" are such patterns. This empirically informed "best guess" then becomes incorporated into the infant's memory and model of language.

PREDICTIVE PROCESSING IMPLIES A PROACTIVE BRAIN

It is important to underscore that from the contemporary perspective the brain is not simply the passive recipient of sensation that is then used to build a model of the world. To the contrary, the brain is proactive (Raichle, 2010). This is captured in Gregory's conception of perceptions as hypotheses (Gregory, 1980). From the perspective of predictive processing the brain has a model of the world before it encounters the world. It uses its model to try to predict what it will experience next in its patterns of firing neurons. Action is taken to sample sensation in a manner that tests the hypothesis (Clark, 2013a). To the extent that the prediction about the state of the world is supported by the sampled sensory data, further processing of the sensation is suppressed as it does not contain useful information (Blakemore et al., 2000). If the prediction is not supported, the resulting prediction error will drive further processing of the sensation. In pursuing the brain's intransigent survival imperative of minimizing prediction error the brain has two main approaches; namely, action and perception (Friston, 2009). With action it can sample the world differently until sampled sensation matches prediction, or it can revise its model. That is, it can update its "prior" to a reality calibrated posterior belief.

PERCEPTUAL INFERENCE-CYCLES OF SEARCHING THE WORLD AND SEARCHING MEMORY

Incoming sensation acts as a retrieval cue for memory (Tulving and Schacter, 1990). For example searching the world with the eyes imports coarse global properties of an object in the form

of patterns of firing neurons. Such patterns are believed to trigger an internal hippocampal mediated search that attempts to answer the question "what is this like?" (Bar and Neta, 2008). From this perspective, object recognition is a matching task. An analogy representing the closest familiar representation in memory is selected by the prefrontal cortex from a matrix of possibilities with differing probabilities (Hakonen et al., 2017). Low probability analogies are suppressed (Depue, 2012). The selected analogy is itself connected to a web of associations. Taken together these activated memory networks correspond to the brain's "best guess" about current reality and what to expect next (Bar, 2009). The brain then tests its prediction by searching the world with saccadic eye movements (Friston et al., 2012). Specifically Friston asserts "...saccadic eye movements are optimal experiments, in which data are gathered to test hypotheses or beliefs..." (Friston, 2012). The data from these eye movements will either support or refute the prediction. If the visual search results do not support the prediction, the brain may attempt to search memory for a new "best guess." This in turn engenders a new visual search to test the new "best guess." Searching the world alternates with searching memory in a constant ongoing flux of processing (Richter et al., 2015). This cycle of sampling, matching from memory, prediction, and further sampling continues throughout life as the brain attempts to navigate the endless uncertainty of incoming sensation by minimizing the errors of its predictions (Clark, 2016).

EYE MOVEMENTS AND HIPPOCAMPUS FORM AN INTEGRATED SEARCH SYSTEM

Perceptual inference as described reflects a process that requires tight coordination between the oculomotor system that controls the movement of the eyes, and the hippocampal search of memory. Converging evidence leads to the conclusion that these two systems are functionally and anatomically coupled (Shen et al., 2016). The nature of this relationship is further illuminated by consideration of the functions of the hippocampus.

THE HIPPOCAMPUS NAVIGATES INTERNAL AND EXTERNAL SPACE

The role of the hippocampus in memory function was first described in Scoville and Milner (1957). Subsequently, it was found to play an important role in spatial navigation (O'Keefe and Dostrovsky, 1971). More recently, these two apparently distinct functions have been reconciled through identification of a common underlying mechanism (Buzsaki and Moser, 2013). A leading theory of hippocampal function posits that the hippocampus acts to index the locations in the cortex of the disparate elements of memory which when co-activated confer the experience of remembering (Teyler and Rudy, 2007). In other words the hippocampus knows where in the cortex to find the smell, the sound, the visual image, etc. of an experience allowing reconstruction of an episodic memory (Schacter and Addis, 2007). See **Figure 1**. In effect the hippocampus maps the physical space inside the brain that gives rise to memories (Bellmund et al., 2018). Similarly the hippocampus maps the disparate landmarks in the physical space outside the brain as it performs its role in navigation in the world. It has been argued that the computational properties of the hippocampus are particularly well-suited to execute this type navigation which is essentially the same whether one is searching the world or searching memory (Buzsaki and Moser, 2013).

THETA RHYTHM KEEPS INFORMATION FLOW ORGANIZED

Consider a walk in the park. Crude visual input triggers a hippocampal mediated memory search and retrieval of the best guess regarding current location. Based on the best guess of current location, the brain predicts the next landmark it will

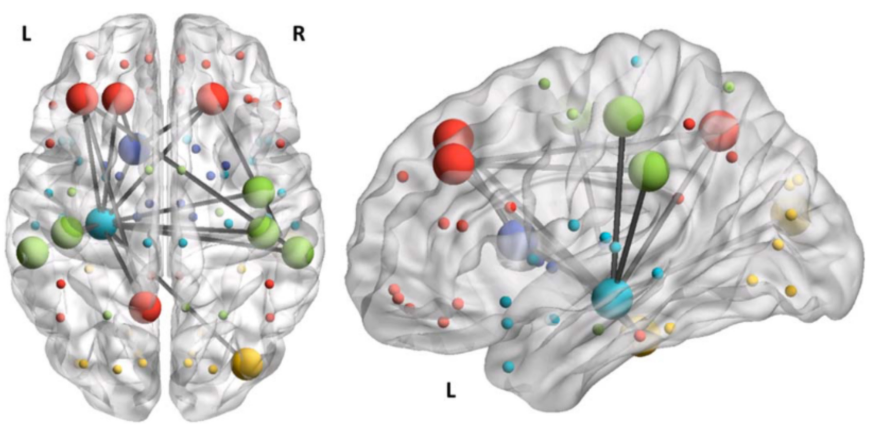

FIGURE 1 | fMRI derived image of successful episodic memory recall showing hippocampal (blue) mediated "retrieval assembly" of cortical regions containing the sensory and motor elements of the memory. Adapted from Geib et al. (2017) and adapted with permission.

encounter using memory (Eichenbaum, 2017). It then tests its prediction by sampling the world with saccadic eye movements (Friston et al., 2012; Parr and Friston, 2017; Stefanics et al., 2018; Smout et al., 2019). In order to execute these cycles the brain must be able to maintain the distinction between new information coming in (encoding) and information already stored in the brain (retrieval). In other words, it must not confuse the monument that is currently seen, with the concession stand that it predicts it will see next based on memory. Recent research suggests that the hippocampal theta rhythm is crucial in organizing the flow of information through the neural circuits responsible for the encoding and retrieval of episodic memory (Hasselmo and Stern, 2014; Siegle and Wilson, 2014).

THETA RHYTHM CORRELATES WITH MEMORY PERFORMANCE

The theta rhythm has been conceived of as "the navigation rhythm through both physical and mnemonic space, facilitating the formation of maps and episodic/semantic memories" (Buzsaki, 2005). More generally theta rhythms promote coordination across distributed brain areas during different types of information processing (Colgin, 2013). Such coordination of disparate regions including the hippocampus and the prefrontal cortex is critical in being able to retrieve episodic memories (Preston and Eichenbaum, 2013; Geib et al., 2017). Recently, theta rhythm synchronization of hippocampal and prefrontal regions by external stimulation has been shown to transiently restore working memory function in older adults (Reinhart and Nguyen, 2019). The implication is that loss of theta synchronization plays an important role in the deterioration of working memory with age. Given that attention and working memory are the most severely compromised neurocognitive functions in PTSD (Scott et al., 2015), the possibility arises that enhanced theta synchronization might improve memory function in PTSD as well.

EYE MOVEMENTS RE-SET THETA RHYTHMS

The ability to import sensory information is highly dependent on the motor rhythms used to acquire that information (Schroeder et al., 2010). While the eyes are moving during a saccade, visual input to the brain is suppressed (Bremmer et al., 2009). We don't perceive this because the brain continues to generate its prediction of the world and fills in the gap. If suppression did not occur, our vision would be like a rapidly panning video camera blurring the image every time the eyes moved. In addition such "sensory attenuation" is necessary to keep incoming sensation separated from brain generated prediction (Brown et al., 2013). When the eyes stop moving there is a period of fixation during which data is acquired (Rajkai et al., 2008). It is primarily during fixation that sensation is taken in to the visual system (Wurtz, 2008; Crevecoeur and Kording, 2017). As previously discussed, organizing the incoming and outgoing flow of information through the hippocampus is essential. It appears that saccadic eye movements play a critical role in this regard by resetting the theta rhythm and thus synchronizing the flow of incoming information through disparate regions including the hippocampus and prefrontal cortex in processing experience and memory (Jutras et al., 2013; Meister and Buffalo, 2016).

SUMMARY OF PERCEPTUAL INFERENCE

Lacking direct experience of the world perception is an act of inference that utilizes raw sensory data to search memory to find the statistically most likely "best guess" about current reality. The best guess is then tested with saccadic eye movements that sample the world obtaining new data that either support or refute the best guess prediction about the state of the world. The cycling of incoming sensation and outgoing prediction based on memory is perpetual while awake. As part of an integrated search system saccadic eye movements reset the theta rhythm synchronizing the flow of incoming and outgoing information from the hippocampus through other participating structures including the prefrontal cortex. Thus disparate regions are coordinated optimizing the processing of current experience.

CROSS REFERENCING SENSORY DATA REDUCES UNCERTAINTY

The challenge of perceptual inference about the state of the world becomes more tractable when data from multiple senses is combined into the computations. For example smell offers probabilities of an orange soda, air freshener, or an actual orange. Vision suggests an orange colored ball, or an orange. Touch suggests a tomato, an apple, or an orange. When the statistical probabilities suggested by each sense are integrated, the most likely single cause of these sensations is an orange. In other words cross-referencing by the senses rapidly reduces the possibilities to the most likely cause.

MEMORY FOR PREDICTION

The clinical relevance of Predictive Processing to psychological trauma and its resolution becomes apparent recognizing that memory is the principle substrate of prediction (Bar, 2009; Buckner, 2010). In fact it can be argued that the raison d'etre of memory at all levels in the brain is to facilitate Predictive Processing (Sterling, 2012). The AIP model postulates "dysfunctionally stored memories" as the foundation of post-traumatic psychopathology. If the core function of the brain is prediction based on memory, it is easy to imagine grossly sub-optimal behavior resulting from such compromised memories. For example when a truck backfires in suburbia, it may trigger a veteran's dysfunctionally stored memory so he predicts incoming mortar fire and dives to the ground. The subsequent absence of destruction from an incoming mortar represents a massive

failure of prediction. If the Predictive Processing account is correct, then the brain would be expected to try to minimize its prediction error to improve future prediction and behavior. The Network Balance Model of Trauma and Resolution postulates that imbalance of the Salience, Default Mode and Central Executive Networks compromises the coordinated interaction of brain regions required to execute this processing (Chamberlin, 2019). However once balance is restored, memory will be processed. But how does memory process? The AIP postulates "there is a system inherent in all of us that is physiologically geared to process information to a state of mental health. . .by means of this system, negative emotions are relieved, and learning takes place, is appropriately integrated, and is available for future use." The Predictive Processing Model of EMDR argues that this "inherent system" is, broadly speaking, Predictive Processing itself. In a sense, it is just what the brain does.

MISMATCH NEGATIVITY REFLECTS PREDICTION ERROR

Mismatch Negativity is a well-established research paradigm that reflects deviation from the brain's expectations (Naatanen et al., 2007). When sensation does not match what is expected, the EEG brainwave recorded over the corresponding sensory cortex will show a negative deflection. For example if the 10th note of the song "Mary had a little lamb" is played incorrectly, people familiar with the song will manifest a negative wave in the EEG over the auditory cortex. The brain is surprised. In the parlance of Predictive Processing the brain registers a prediction error. If the song is played incorrectly in the same way multiple times, the magnitude of the measured prediction error will diminish as the brain updates its model and expectations (Baldeweg, 2007; Garrido et al., 2009). In neuro-energetic modeling the magnitude of mismatch negativity has been shown to correlate to the magnitude of prediction error and reflects an increase in energy available to drive synaptic adaptation (Strelnikov, 2007). In effect the brain learns to predict a different pattern of firing neurons (sound) under certain circumstances, e.g., when the song is played by a 5 year-old novice. Prediction error has been minimized. Similar effects have been demonstrated in the visual realm using facial expressions that are unexpected, thus supporting the Predictive Processing paradigm and its postulated updating of models (Stefanics et al., 2018). In the tactile realm, unexpected changes in the intensity of a stimulus result in the updating of somatic models, an effect mediated by the anterior insula (Allen et al., 2015). This has significant implications for the awareness of somatic sensation in trauma, and its processing in EMDR Therapy utilizing bilateral tactile stimuli. It is important to note that mismatch negativity and related violation responses later in peri-stimulus time (e.g., the P 300) are not limited to sensation, but has also been demonstrated on the conceptual levels of grammar and semantic meaning (Naatanen et al., 2007; Batterink and Neville, 2013). Thus Mismatch Negativity appears to reflect the occurrence of prediction error in the brain on multiple levels and in multiple regions.

MEMORY RECONSOLIDATION

Following retrieval a memory must undergo a process involving protein synthesis called Memory Reconsolidation in order to return to storage (Nader et al., 2000). This creates an opportunity to alter the memory in several ways. For example pharmacologic interventions may disrupt protein synthesis so the memory cannot return to storage (Przybyslawski et al., 1999; Debiec and Ledoux, 2004; Kindt et al., 2014). Effectively the memory is erased. While promising as a clinical intervention there are many constraints that need to be navigated for the approach to be useful. These constraints are currently the subject of active research (Visser et al., 2018). Another way memory may be altered is physiologically through memory updating or learning. Not only are the causes of sensation uncertain, but the future is inherently uncertain because self and world are constantly changing. This requires that memories be capable of being updated when conditions change in order to optimize predictions (Dudai, 2009; Lee, 2009; Kroes and Fernandez, 2012).

MEMORY RECONSOLIDATION CONSTRAINED BY BOUNDARY CONDITIONS

The process of updating memory is achieved through Memory Reconsolidation and requires that certain conditions be satisfied (Monfils et al., 2009). The first so-called "boundary condition" to be described was a "mismatch between what is expected and what actually occurs" (Pedreira et al., 2004). After a memory is retrieved, if there is a mismatch, i.e., a prediction error, the memory may enter an active state during which new information can be incorporated into the memory. Under these conditions the memory will be updated. However if there is a significant delay in receiving the information that contradicts the expectation, the result will be "extinction," that is the creation of a new competing memory instead of updating the original memory (Diaz-Mataix et al., 2013). In addition if the new information/experience is too dissimilar to the retrieved memory, a new memory will be created and the retrieved memory will be left intact (Forcato et al., 2009). Thus identifying and controlling boundary conditions are critical if the Memory Updating process is to be harnessed in a therapeutic fashion (Schiller et al., 2010; Kroes et al., 2016; Walker and Stickgold, 2016; Elsey and Kindt, 2017; Treanor et al., 2017).

PREDICTION ERROR WINDOW OF MEMORY RECONSOLIDATION

Several points are worth highlighting regarding memory updating following experience. Memory updating can and does occur spontaneously and without conscious awareness throughout life. And when it occurs, the changes span the range from "intracellular gene inductions to brain-wide systems level reorganization of memory representations" (Stickgold and Walker, 2007). This is consistent with the theoretical formulation of Predictive Processing based on statistical physics that asserts

that all quantities that can change in a system will change, in order to minimize prediction error (Friston et al., 2006; Strelnikov, 2010). Furthermore the magnitude of the prediction error appears to be critical in regulating memory updating when attempting a therapeutic intervention. If the prediction error is small, the memory will be not be updated given lack of significant new information. In contrast, if the prediction error is too large, the brain appears to treat it as a new experience and creates a new memory. The original memory is not updated. Only if the prediction error is "moderate" does updating of the memory with new information via reconsolidation occur (Finnie and Nader, 2012; Sevenster et al., 2014; Beckers and Kindt, 2017). In The Predictive Processing Model of EMDR this optimal level of moderate prediction error is referred to as the Prediction Error Window and is a crucial factor in harnessing the therapeutic potential of Memory Reconsolidation.

PROCESSING TRAUMATIC EXPERIENCE-NETWORK BALANCE

It has been postulated that balance of the three principle large-scale networks is an essential pre-requisite for the processing of traumatic experience to an optimal state (Chamberlin, 2019). Such balance may occur spontaneously or through effective trauma therapies such as EMDR. Elements of the EMDR protocol are thought to activate specific individual networks. For example questions during the assessment phase are posited to activate the default mode and salience networks bringing the individual into a state of active re-experiencing. Subsequently therapist guided Dual Attention and Eye Movements are posited to have a crucial role in activating the central executive network thus restoring network balance. This allows the individual to begin taking in new information from the external world and orienting to the present. In essence these interventions set the stage for the "inherent system" postulated by the AIP to then spontaneously process the "dysfunctionally stored memory." The Predictive Processing Model of EMDR suggests how the "inherent system" may actually function.

PROCESSING TRAUMATIC EXPERIENCE-PREDICTION ERROR MINIMIZATION

Having been brought into a state of active re-experiencing during EMDR, the brain predicts what will come next as the remembered trauma unfolds. For example an individual involved in a car accident predicts the sight of broken glass, the smell of burning plastic, and the pressure of an airbag on the chest. The Predictive Processing Model of EMDR postulates the following sequence of events: saccadic eye movements guided by the therapist compel multi-modal sampling of current sensation thus testing the individual's predictions of what comes next. Sensory sampling of the therapist's office does not support the predicted car accident mayhem. There is no broken glass, smell of burning plastic or airbag pressure. The result is multi-modal prediction error. This prediction error registers in the brain as Mismatch Negativity in multiple regions. Energy is mobilized for synaptic adaptation. Memory reconsolidation is initiated. Subsequent sampling is invoked to generate new predictions as the individual attains progressively greater orientation to the benign present. All the while saccadic eye movement mediated theta rhythm synchronization keeps the inflow of sensation and outflow of mnemonic predictions organized for optimal processing. Disparate brain regions are synchronized, and working memory capacity is restored. The net result, the overarching goal of the brain, is prediction error minimization. Ultimately prediction error minimization is driven by the thermodynamics of free energy minimization (Friston, 2010; Sengupta et al., 2013). "There was a car accident but it's not happening now. It's over and I'm sitting in an office."

RE-ENTRANT PROCESSING

The preceding example of prediction error minimization occurs over time with repeated sets of subjective reports followed by sets of eye movements. Presumably this involves cycles of re-entrant processing as information gets passed through thalamo-cortical as well as cortico-cortico loops as the processes of disambiguation, differentiation, sensory integration, and mnemonic integration occur (Edelman and Gally, 2013; Preston and Eichenbaum, 2013; Ohkawa et al., 2015; Richter et al., 2015; Hakonen et al., 2017; Kitamura et al., 2017; Yokose et al., 2017; Chao et al., 2018). The result is an updated memory and model of the world that makes better predictions.

PRECISION WEIGHTING OF PREDICTION ERROR

Implicit in the preceding discussion of prediction error minimization is the predictive processing mechanism of precision weighting of prediction error. As noted by Helmholtz incoming sensation is ambiguous. In addition the sensory signal itself it is often imprecise and unreliable. Potentially this sets the stage for an unreliable signal to drive prediction error minimization and memory updating thus compromising the future utility of the brain's generative model of the world. The predictive processing response to this challenge, postulated to reflect brain function, is to offer a prediction regarding the reliability of the sensory signal. This prediction of reliability is called "precision weighting" and reflects the degree of confidence or precision in the sensory signal (Friston, 2009). It is an estimate of uncertainty that reflects the trustworthiness of the sensation and is posited to be implemented biologically through changes in synaptic gain modulated by "top down" cortical predictions. Signals deemed unreliable and imprecise carry less weight or influence, and are not able to drive learning. In a sense the downward flowing prediction or

belief prevails as the interpretation of the current state of the world, and the prediction error based on unreliable sensation is ignored. We then experience what we expect, rather than what sensation might suggest. In contrast, signals deemed reliable have more "weight" with increased synaptic gain, and are able drive memory updating. Thus precision weighting of prediction error can be conceived of as a mechanism for modulating the influence of prediction errors on belief updating (Clark, 2013b). In other words, precision weighting helps us to implicitly ask and answer the following questions: "How much do I trust current sensation? Which sensory channels are the most reliable? And "do I need to update my beliefs?"

PRECISION GIVES RISE TO ATTENTION

Precision weighting also offers a way of understanding sensory attention at a neuronal level. Formally this has been expressed as "attention is the process of optimizing the synaptic gain to represent the precision of sensory information during hierarchical inference" (Feldman and Friston, 2010). This proposition has received strong empirical support in a study of spatial attention and response speed (Vossel et al., 2014). From this perspective attention is an emergent property of the process of estimating the reliability of sensation. As the brain estimates the uncertainty associated with different channels of sensation, giving more weight to some channels through synaptic gain, and less weight to others, the byproduct is what we call "paying attention" (Hohwy, 2013). For example, a sailor proceeding through a dense fog may predict that sound is more reliable than vision, and as a result pays more attention to sound, and relies less on vision than he would on a clear day. Given that the deployment of attention is a crucial factor in the pathology of PTSD, precision weighting may play an important role.

ATTENTION COMPROMISED IN PTSD

A recent meta-analysis found that attention and working memory were among the most severely compromised neurocognitive functions in PTSD (Scott et al., 2015).

Early investigators characterized the abnormalities in attention seen in PTSD as a bias toward threatening stimuli (Fani et al., 2012). While this is frequently demonstrated in clinical populations, there is also a high incidence of bias away from threat, i.e., ignoring threat (Bar-Haim et al., 2010; Sipos et al., 2014). This observation led to the recognition that the abnormalities in attention seen in PTSD are characterized by an increase in Attention Bias Variability (ABV). PTSD sufferers are biased toward the extremes of attention, i.e., excessive attention towards threat at times, and excessive attention away from threat at other times (contributing to reckless behavior) (Iacoviello et al., 2014; Naim et al., 2015). In PTSD the control and deployment of attention appears compromised, thus raising the question of how this might be influenced by precision weighting.

PRECISION IN PSYCHOLOGICAL TRAUMA

Recent empirical work has explored the potential effects of threat on precision weighting in humans (Cornwell et al., 2017). The authors found that under threat of unpredictable aversive shock, there was an increased auditory mismatch response to deviant stimuli best explained by increased post-synaptic gain in primary auditory cortex, with precision weighting biased toward feed forward propagation of prediction errors. This was consistent with a state of anxious hypervigilance and attentional bias to threats in the environment.

Considering the potential role of precision weighting in psychological trauma, Wilkinson et al. (2017) suggested that the survival imperative resulting from experience of a life-threatening event might result in an unusually strong prior probability that will be selected, even when the incoming sensation is a relatively poor fit. The position seems to be, "I must act to ensure survival, evidence be damned." Recently, this concept has been explored empirically.

Using an agent-based model computer simulation of PTSD (Linson et al., 2019) varied the precision weighting of a prior cued by a stressor and found a perturbation in the balance between exploration and exploitation. Specifically, when the prior was afforded low precision the agent engaged in exploration and evidence accumulation, essentially testing the hypothesis "I'm in danger." However, when the prior was afforded high precision the agent exploited its knowledge of how to avoid danger and took defensive action, without actually assessing if it was in danger. This was accompanied by physiological responses characteristic of PTSD coded into the model. The authors interpreted these findings by suggesting that a prior belief that carries a high probability of injury or death is afforded high precision via natural selection given that the potential catastrophic consequences outweigh the benefits of exploration and evidence accumulation. (A familiar example of this might be herd behavior when a group of animals run from a predator. While only a subset actually saw the predator, their running triggers in the others the prior of a predator and they take defensive action and start running, without actually trying to see if there is a predator or not. "Better safe than sorry.").

This suggests that altered precision may result in a state biased against evidence accumulation, consistent with impairments in safety learning characteristic of PTSD (Jovanovic et al., 2012; Sijbrandij et al., 2013). What has been called "safety blindness" (Chamberlin, 2019). And further, that EMDR may act in part by overcoming this bias thus facilitating the acquisition of evidence that does not support traumatic experience in the present. Analysis of the role of eye movements in attention can help illustrate how this might work.

ATTENTION, PRECISION, AND EYE MOVEMENTS

What we see depends on where we look. And where we look depends on a guess, a prediction, about where we can find

what we are looking for. And what we actually find there, in turn informs where we look next (Parr and Friston, 2017). Thus the motor element of attention, where we look, and the perceptual element, what we see, are mutually informative and interdependent (Mirza et al., 2016). The perceptual and motor elements are part of the perpetual circular processing of the Perception-Action Cycle (Fuster and Bressler, 2015). Elucidating the precise anatomy and physiology of this cycle of visual foraging has been a major challenge for cognitive neuroscience.

The first element involves the motor system and the overt orienting of attention with saccadic eye movements. The second element involves perception and the covert orientation of attention without eye movement. This is the aspect of attention most directly related to precision as previously discussed. This entails orienting to sensation that offers the best evidence in support of the current belief that is being tested (Friston et al., 2012; Mirza et al., 2018). These motor and sensory aspects of attention are tightly coupled sharing a largely overlapping neuroanatomy (Corbetta et al., 1998; Nobre et al., 2000; de Haan et al., 2008). Sharing essentially the same anatomy yet performing dissociable motor and sensory functions (Juan et al., 2008) presents a dilemma that has thus far has defied satisfying explanation. By incorporating neural oscillations the Rhythmic Theory of Attention suggests how this dilemma might be resolved (Fiebelkorn and Kastner, 2019a).

Based on empiric data in humans and monkeys Fiebelkorn and Kaster found rhythmic epochs of enhanced sensory sensitivity alternating with saccadic eye movements during specific phases of theta rhythm. The "sampling state" was characterized by enhanced sensory processing and suppression of attentional shifts, both covert and overt. The "shifting state" was characterized by an attenuation of sensory processing, and was sometimes associated with a covet shift, and sometimes an overt shift in attention. The authors interpreted this to be a state of disengagement that creates an opportunity to shift, either overtly or covertly. These theta "clocked" states were associated with a rhythmic reweighting of network connections to either support motor or sensory activity (Fiebelkorn and Kastner, 2019b). The Rhythmic Theory of Attention posits that the theta rhythm organizes environmental sampling by periodically reweighting functional connections to motor or sensory regions resulting in states that promote either sampling or shifting. Thus the deployment of both overt (saccadic) and covert (precision mediated) attention are tightly coupled, intimately associated with eye movement, and organized by the theta rhythm ("clocking").

EMDR MAY RESTORE ATTENTION

Taken together these considerations suggest that the therapeutic target of EMDR in PTSD may be in overcoming the bias against exploration and evidence accumulation. Challenging this bias repeatedly with sets of therapist guided eye movements may restore the integrity of the rhythmic deployment of attention (overt and covert) leading to evidence accumulation of a non-traumatic present, recalibration of sensory precision, and the updating of memory. The net result of treatment with EMDR may be relearning how to deploy attention and weigh the sensory evidence we receive from inside and outside the body in support of our narrative about what is happening now. If so, this hypothesis predicts a reduction in ABV and aberrant theta activity following successful treatment with EMDR (Dunkley et al., 2015). Such a reduction would be consistent with recent work utilizing attention control training that resulted in a decrease in PTSD symptoms, ABV (Badura-Brack et al., 2015) and aberrant theta activity (McDermott et al., 2016).

Having described the core elements of the Predictive Processing Model of EMDR it is now possible to posit how some common clinical phenomena from the practice of EMDR might be explained.

EYES MOVE TO REMEMBER

Previous discussion of the tight coupling between the oculomotor system and hippocampus elucidated how eye movements can drive search of memory to identify an analogy that matches current sensation thus forming the brain's best guess. Another manifestation of this integrated oculomotor–hippocampal system is the search of memory that results from eye movements without regard to sensation.

During conversation individuals will periodically look away from the person they are talking to toward regions of the visual field that do not contain any useful information. This so called "Looking at nothing" phenomenon has spawned research that suggests it has an important role in cognition. Also called "non-visual eye movements" or "non-visual gaze paths" the core hypothesis of this research is that saccadic eye movements play a role in non-visual cognitive tasks (Ehrlichman et al., 2007). One finding is that rates of non-visual eye movements increase in tasks requiring search of long term memory and episodic recall (Micic et al., 2010). Going beyond simple association early research found that performance of episodic recall is enhanced with saccadic eye movements (Christman et al., 2003). Subsequent research has established the facilitation of retrieval from memory by eye movements consistent with the concept of embodied cognition (Bochynska and Laeng, 2015; Scholz et al., 2016). The idea is that the specific gaze path traced during the encoding of an experience may enhance recall when it is physically re-enacted during retrieval. Alternatively restriction of eye movement has been shown to impair memory performance (Johansson et al., 2012; Laeng et al., 2014). And finally memory processing during REM sleep is characterized by the elaboration of wide ranging associations while the eyes are closed. These findings suggest that saccadic eye movements have an important role in search and retrieval from memory that is independent of visual input. (The classic Analytic geometry of lying down and staring at the ceiling to facilitate free association appears to support this idea). Indeed it has been suggested that "there is an inherent link, functionally and anatomically between the brain's oculomotor system and its hippocampal system" (Liu et al., 2016). And further that the physiological coupling between these systems may be obligatory (Andrillon et al., 2015).

Taken together these findings suggest the possibility that the operation of the occulomotor–hippocampal system, like many systems in the brain, is bi-directional, and that eye movements may be used deliberately to drive memory search (Christman et al., 2003, 2006; Parker et al., 2008; Brunye et al., 2009; Parker and Dagnall, 2010). While EMDR therapy appears to have incorporated and capitalized on this phenomenon, the explanation offered has been "increased inter-hemispheric brain activity" rather than the bi-directional oculomotor–hippocampal hypothesis advanced above.

CLEARING THE CHANNELS OF TRAUMATIC MEMORY

From the preceding discussion it appears that eye movements sometimes occur in the service of vision, e.g., searching the world and sampling reality to test a prediction. In addition eye movements may also occur in the service of memory, e.g., searching memory for associations. Recalling the example of a walk in the park, these distinct roles spontaneously shift rapidly and flexibly throughout waking life. An interesting manifestation of this shifting can be seen during successive sets of eye movements during EMDR. Assessment questions bring the traumatic memory online often engendering a rising level of arousal. Initial sets of eye movements typically result in a rapid "desensitization" with decreasing arousal (Elofsson et al., 2008; Schubert et al., 2011). This appears to result from central executive network activation and amygdala deactivation (de Voogd et al., 2018). The second is increased sampling of current sensation testing the predictions of the traumatic memory (Mirza et al., 2018). This reflects the use of eye movements in the service of vision. When current sensation does not support the predictions of trauma, arousal decreases. However subsequent sets of eye movements are often associated with an increased level of arousal (Sack et al., 2008). Per the Predictive Processing Model this occurs as a result of eye movements in the service of memory. (Seeking uncertainty leads to opportunities to reduce uncertainty). Specifically, eye movements drive the search for associations often finding a new traumatic memory fragment with its corresponding prediction. Returning to the car accident example, the prediction of broken glass is tested by using eye movements and is not supported. Prediction error is minimized and arousal decreases. The next set of eye movements drives memory search and finds the associated fragment of active bleeding from lacerations. This is accompanied by fear and increased arousal. The next set of eye movements then prompts searching the world to test this prediction. The prediction of bleeding is not supported resulting in a fall in arousal. Clinically one result of these cycles of searching the world and then searching memory is an undulating level of arousal with an overall downward trend. The predictions of the traumatic memory, and all its associations are progressively found, tested and not supported. In the words of Friston, "the only hypothesis that can endure over successive saccades is the one that correctly predicts the salient features that are sampled" (Friston, 2012). This leads to an inevitable best guess of current reality: "no trauma happening now."

ATTENTION AMPLIFIES PREDICTION ERROR

During processing of traumatic experience with EMDR residual symptoms may persist despite significant attenuation. The therapist may then direct the client's attention to one of the residual symptoms. For example after learning that there is still an abnormal sensation in the abdomen, the therapist may instruct the client to "Go with that" before initiating another set of eye movements. Clinical experience suggests that this intervention is effective in facilitating complete processing of the memory. But how does this work?

Recent research has demonstrated that directing attention to a prediction error amplifies the error signal thus enhancing the neural encoding of the error (Smout et al., 2019). This suggests that the therapist's directing attention to a residual symptom may amplify the prediction error prompting the brain to minimize the error. That is, the prediction error has been amplified to the "moderate range" where it is in the Prediction Error Window that triggers memory reconsolidation. This appears to result in a complete resolution of the symptom.

LINKING TO ADAPTIVE NETWORKS

Another important clinical phenomenon related to eye movement driven elaboration of associations is the linking of the traumatic experience to "adaptive networks" of memory as postulated by the AIP. Network research suggests that elaboration of associations (mental exploration) is the default mode of the brain (Buckner et al., 2008). When the "load" of cognitive and perceptual processing demand is low, the brain searches memory widely (Baror and Bar, 2016). In contrast when the demands are high, the brain utilizes immediate "obvious" information from memory without significant search. For example if there is a gun in your face, your thoughts and associations will probably be very narrowly focused on escape from danger, e.g., door, window. You are unlikely to be reflecting on how guns helped promote survival on the western frontier, or the implications for society of being able to make guns digitally from 3-D printers. The Predictive Processing Model of EMDR postulates that as the prediction error of traumatic memory gets reduced, demand decreases and eye movements drive elaboration of associations progressively more distant from those of the core memory (Christman et al., 2006; Parker et al., 2008, 2009; El Khoury-Malhame et al., 2011). Initially associations will be local, i.e., closely related to, or part of the trauma. From the preceding car accident example, associations might be to bleeding from lacerations, or the ambulance ride to the ER. As demand and arousal decreases associations are broader, and more "global." For example, "recovering from this car accident was like when I rebounded from the skiing accident. I'm pretty resilient." This results in a state where there is co-activation

of two previously unrelated memories simultaneously. Such synchronous co-activation has been shown to result in formation of a qualitatively new memory that links the previously independent memories (Ohkawa et al., 2015; Yokose et al., 2017). One result is that activation of one memory, e.g., car accident, now triggers activation of the newly linked memories of ski accident and resilience. Per the AIP, the traumatic memory has been linked to an adaptive network.

MODEL PREDICTIONS

The Predictive Processing Model of EMDR contains multiple predictions that can be empirically tested. For example the model predicts that the processing of traumatic memory with saccadic eye movements in the benign present will result in an increase in prediction error. If so, the increase in prediction error should be reflected by an increase in mismatch negativity. It is also postulated that processing entails serial predictions as associated memory fragments are recalled and tested. If so, the increase in mismatch negativity would be expected to undulate and potentially be synchronized with the undulation of arousal that has been measured. That is an increase in arousal as a new prediction arises, followed by increased mismatch negativity as it is tested and not supported. Over the entire session mismatch negativity (and arousal) would be expected to drop as the memory is updated and the benign present becomes predicted.

The model also suggests that EMDR may act to restore the integrity of the rhythmic deployment of attention including the re-calibration of precision weighting. If so, this would be expected to result in a decrease in aberrant theta dynamics, and a decrease in ABV in patients who experience significant improvement in symptoms.

CONCLUSION

The Predictive Processing Model of EMDR builds on The Network Balance Model of Trauma and Resolution utilizing the foundation of the Free Energy Principle to explain how traumatic memories are resolved using EMDR as an example. With the progressive restoration of large-scale network balance, the physiological conditions necessary for the optimal processing of memory are re-established. Next, driven by an excess of Free Energy the brain resumes prediction error minimization of the traumatic memory. Saccadic eye movements facilitate this Predictive Processing resulting in memory updating with reconsolidation and integration into widespread mnemonic networks. EMDR therapy was be used to illustrate how specific clinical interventions may facilitate the processing of "dysfunctionally stored memory" and the resolution of trauma.

DATA AVIALABILITY STATEMENT

No new data sets were generated for this review of previously published studies.

AUTHOR CONTRIBUTIONS

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

FUNDING

Research and publication funded by the author. There are no outside sources of funding.

REFERENCES

Allen, M., Fardo, F., Dietz, M. J., Hillebrandt, H., Friston, K., Rees, G., et al. (2015). Anterior insula coordinates hierarchical processing of tactile mismatch responses. *Neuroimage* 127, 34–43. doi: 10.1016/j.neuroimage.2015.11.030

Andrillon, T., Nir, Y., Cirelli, C., Tononi, G., and Fried, I. (2015). Single-neuron activity and eye movements during human REM sleep and awake vision. *Nat. Commun.* 6:7884. doi: 10.1038/ncomms8884

Badura-Brack, A. S., Naim, R., Ryan, T. J., Levy, O., Abend, R., Khanna, M. M., et al. (2015). Effect of attention training on attention bias variability and PTSD symptoms: randomized controlled trials in Israeli and U.S. combat veterans. *Am. J. Psychiatry* 172, 1233–1241. doi: 10.1176/appi.ajp.2015.14121578

Baldeweg, T. (2007). ERP repetition effects and mismatch negativity generation. *J. Psychophysiol.* 21, 204–213. doi: 10.1027/0269-8803.21.34.204

Bar, M. (2009). The proactive brain: memory for predictions. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 364, 1235–1243. doi: 10.1098/rstb.2008.0310

Bar, M., and Neta, M. (2008). The proactive brain: using rudimentary information to make predictive judgements. *J. Consum. Behav.* 7, 313–330.

Bar-Haim, Y., Holoshitz, Y., Eldar, S., Frenkel, T. I., Muller, D., Charney, D. S., et al. (2010). Life-threatening danger and suppression of attention bias to threat. *Am. J. Psychiatry* 167, 694–698. doi: 10.1176/appi.ajp.2009.09070956

Baror, S., and Bar, M. (2016). Associative activation and its relation to exploration and exploitation in the brain. *Psychol. Sci.* 27, 776–789. doi: 10.1177/0956797616634487

Batterink, L., and Neville, H. J. (2013). The human brain processes syntax in the absence of conscious awareness. *J. Neurosci.* 33, 8528–8533. doi: 10.1523/jneurosci.0618-13.2013

Beckers, T., and Kindt, M. (2017). Memory reconsolidation interference as an emerging treatment for emotional disorders: strengths, limitations, challenges, and opportunities. *Annu. Rev. Clin. Psychol.* 13, 99–121. doi: 10.1146/annurev-clinpsy-032816-045209

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., and Doeller, C. F. (2018). Navigating cognition: spatial codes for human thinking. *Science* 362:eaat6766. doi: 10.1126/science.aat6766

Bergmann, U. (2010). EMDR's neurobiological mechanisms of action: a survey of 20 years of searching. *J. EMDR Pract. Res.* 4, 22–42. doi: 10.1891/1933-3196.4.1.22

Bisson, J. I., Roberts, N. P., Andrew, M., Cooper, R., and Lewis, C. (2013). Psychological therapies for chronic post-traumatic stress disorder (PTSD) in adults. *Cochrane Database Syst. Rev.* 12:Cd003388. doi: 10.1002/14651858.CD003388.pub4

Blakemore, S. J., Wolpert, D., and Frith, C. (2000). Why can't you tickle yourself? *Neuroreport* 11, R11–R16.

Bochynska, A., and Laeng, B. (2015). Tracking down the path of memory: eye scanpaths facilitate retrieval of visuospatial information. *Cogn. Process* 16(Suppl. 1), 159–163. doi: 10.1007/s10339-015-0690-0

Bremmer, F., Kubischik, M., Hoffmann, K. P., and Krekelberg, B. (2009). Neural dynamics of saccadic suppression. *J. Neurosci.* 29, 12374–12383. doi: 10.1523/jneurosci.2908-09.2009

Brown, H., Adams, R. A., Parees, I., Edwards, M., and Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cogn. Process* 14, 411–427. doi: 10.1007/s10339-013-0571-3

Brunye, T. T., Mahoney, C. R., Augustyn, J. S., and Taylor, H. A. (2009). Horizontal saccadic eye movements enhance the retrieval of landmark shape and location information. *Brain Cogn.* 70, 279–288. doi: 10.1016/j.bandc.2009.03.003

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annu. Rev. Psychol.* 61, 27–48. doi: 10.1146/annurev.psych.60.110707.163508

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011

Buzsaki, G. (2005). Theta rhythm of navigation: link between path integration and landmark navigation, episodic and semantic memory. *Hippocampus* 15, 827–840. doi: 10.1002/hipo.20113

Buzsaki, G., and Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nat. Neurosci.* 16, 130–138. doi: 10.1038/nn.3304

Chamberlin, D. E. (2019). The network balance model of trauma and resolution—level I: large-scale neural networks. *J. EMDR Pract. Res.* 13, 124–142. doi: 10.1891/1933-3196.13.2.124

Chao, Z. C., Takaura, K., Wang, L., Fujii, N., and Dehaene, S. (2018). Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. *Neuron* 100, 1252.e3–1266.e3. doi: 10.1016/j.neuron.2018.10.004

Christman, S. D., Garvey, K. J., Propper, R. E., and Phaneuf, K. A. (2003). Bilateral eye movements enhance the retrieval of episodic memories. *Neuropsychology* 17, 221–229. doi: 10.1037/0894-4105.17.2.221

Christman, S. D., Propper, R. E., and Brown, T. J. (2006). Increased interhemispheric interaction is associated with earlier offset of childhood amnesia. *Neuropsychology* 20, 336–345. doi: 10.1037/0894-4105.20.3.336

Clark, A. (2013a). Expecting the world: perception, prediction, and the origins of human knowledge. *J. Philos.* 110, 469–496. doi: 10.5840/jphil2013110913

Clark, A. (2013b). The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Front. Psychol.* 4:270. doi: 10.3389/fpsyg.2013.00270

Clark, A. (2013c). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Clark, A. (2016). *Surfing Uncertainty : Prediction, Action, and The Embodied Mind*. Oxford: Oxford University Press.

Colgin, L. L. (2013). Mechanisms and functions of theta rhythms. *Annu. Rev. Neurosci.* 36, 295–312. doi: 10.1146/annurev-neuro-062012-170330

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., et al. (1998). A common network of functional areas for attention and eye movements. *Neuron* 21, 761–773. doi: 10.1016/s0896-6273(00)80593-0

Cornwell, B. R., Garrido, M. I., Overstreet, C., Pine, D. S., and Grillon, C. (2017). The unpredictive brain under threat: a neurocomputational account of anxious hypervigilance. *Biol. Psychiatry* 82, 447–454.doi: 10.1016/j.biopsych.2017.06.031

Crevecoeur, F., and Kording, K. P. (2017). Saccadic suppression as a perceptual consequence of efficient sensorimotor estimation. *eLife* 6:e25073. doi: 10.7554/eLife.25073

de Haan, B., Morgan, P. S., and Rorden, C. (2008). Covert orienting of attention and overt eye movements activate identical brain regions. *Brain Res.* 1204, 102–111. doi: 10.1016/j.brainres.2008.01.105

de Voogd, L. D., Kanen, J. W., Neville, D. A., Roelofs, K., Fernandez, G., and Hermans, E. J. (2018). Eye-movement intervention enhances extinction via amygdala deactivation. *J. Neurosci.* 38, 8694–8706. doi: 10.1523/jneurosci.0703-18.2018

Debiec, J., and Ledoux, J. E. (2004). Disruption of reconsolidation but not consolidation of auditory fear conditioning by noradrenergic blockade in the amygdala. *Neuroscience* 129, 267–272. doi: 10.1016/j.neuroscience.2004.08.018

Dehaene, S., and Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227. doi: 10.1016/j.neuron.2011.03.018

Depue, B. E. (2012). A neuroanatomical model of prefrontal inhibitory modulation of memory retrieval. *Neurosci. Biobehav. Rev.* 36, 1382–1399. doi: 10.1016/j.neubiorev.2012.02.012

Diaz-Mataix, L., Ruiz Martinez, R. C., Schafe, G. E., LeDoux, J. E., and Doyere, V. (2013). Detection of a temporal error triggers reconsolidation of amygdala-dependent memories. *Curr. Biol.* 23, 467–472. doi: 10.1016/j.cub.2013.01.053

Dudai, Y. (2006). Reconsolidation: the advantage of being refocused. *Curr. Opin Neurobiol.* 16, 174–178. doi: 10.1016/j.conb.2006.03.010

Dudai, Y. (2009). Predicting not to predict too much: how the cellular machinery of memory anticipates the uncertain future. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 364, 1255–1262. doi: 10.1098/rstb.2008.0320

Dunkley, B. T., Sedge, P. A., Doesburg, S. M., Grodecki, R. J., Jetly, R., Shek, P. N., et al. (2015). Theta, mental flexibility, and post-traumatic stress disorder: connecting in the parietal cortex. *PLoS One* 10:e0123541. doi: 10.1371/journal.pone.0123541

Edelman, G. M., and Gally, J. A. (2013). Reentry: a key mechanism for integration of brain function. *Front. Integr. Neurosci.* 7:63. doi: 10.3389/fnint.2013.00063

Ehrlichman, H., Micic, D., Sousa, A., and Zhu, J. (2007). Looking for answers: eye movements in non-visual cognitive tasks. *Brain Cogn.* 64, 7–20. doi: 10.1016/j.bandc.2006.10.001

Eichenbaum, H. (2017). The role of the hippocampus in navigation is memory. *J. Neurophysiol.* 117, 1785–1796. doi: 10.1152/jn.00005.2017

El Khoury-Malhame, M., Lanteaume, L., Beetz, E. M., Roques, J., Reynaud, E., Samuelian, J. C., et al. (2011). Attentional bias in post-traumatic stress disorder diminishes after symptom amelioration. *Behav. Res. Ther.* 49, 796–801. doi: 10.1016/j.brat.2011.08.006

Elofsson, U. O., von Scheele, B., Theorell, T., and Sondergaard, H. P. (2008). Physiological correlates of eye movement desensitization and reprocessing. *J. Anxiety Disord.* 22, 622–634. doi: 10.1016/j.janxdis.2007.05.012

Elsey, J. W. B., and Kindt, M. (2017). Breaking boundaries: optimizing reconsolidation-based interventions for strong and old memories. *Learn. Mem.* 24, 472–479. doi: 10.1101/lm.044156.116

Fani, N., Tone, E. B., Phifer, J., Norrholm, S. D., Bradley, B., Ressler, K. J., et al. (2012). Attention bias toward threat is associated with exaggerated fear expression and impaired extinction in PTSD. *Psychol. Med.* 42, 533–543. doi: 10.1017/s0033291711001565

Feldman, H., and Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* 4:215. doi: 10.3389/fnhum.2010.00215

Fiebelkorn, I. C., and Kastner, S. (2019a). A rhythmic theory of attention. *Trends Cogn. Sci.* 23, 87–101. doi: 10.1016/j.tics.2018.11.009

Fiebelkorn, I. C., and Kastner, S. (2019b). Functional specialization in the attention network. *Annu. Rev. Psychol.* 71, 1–29. doi: 10.1146/annurev-psych-010418-103429

Finnie, P. S., and Nader, K. (2012). The role of metaplasticity mechanisms in regulating memory destabilization and reconsolidation. *Neurosci. Biobehav. Rev.* 36, 1667–1707. doi: 10.1016/j.neubiorev.2012.03.008

Forcato, C., Argibay, P. F., Pedreira, M. E., and Maldonado, H. (2009). Human reconsolidation does not always occur when a memory is retrieved: the relevance of the reminder structure. *Neurobiol. Learn. Mem.* 91, 50–57. doi: 10.1016/j.nlm.2008.09.011

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2012). Embodied inference and spatial cognition. *Cogn. Process* 13(Suppl. 1), S171–S177. doi: 10.1007/s10339-012-0519-z

Friston, K., Adams, R. A., Perrinet, L., and Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Front. Psychol.* 3:151. doi: 10.3389/fpsyg.2012.00151

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Fuster, J. M., and Bressler, S. L. (2015). Past makes future: role of pFC in prediction. *J. Cogn. Neurosci.* 27, 639–654. doi: 10.1162/jocn_a_00746

Garrido, M. I., Kilner, J. M., Kiebel, S. J., Stephan, K. E., Baldeweg, T., and Friston, K. J. (2009). Repetition suppression and plasticity in the human brain. *Neuroimage* 48, 269–279. doi: 10.1016/j.neuroimage.2009.06.034

Geib, B. R., Stanley, M. L., Dennis, N. A., Woldorff, M. G., and Cabeza, R. (2017). From hippocampus to whole-brain: the role of integrative processing in episodic memory retrieval. *Hum. Brain Mapp.* 38, 2242–2259. doi: 10.1002/hbm.23518

Geisler, W. S., and Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cogn. Sci.* 27, 379–402. doi: 10.1207/s15516709cog2703_3

Gregory, R. L. (1980). Perceptions as hypotheses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 290, 181–197.

Hakonen, M., May, P. J. C., Jaaskelainen, I. P., Jokinen, E., Sams, M., and Tiitinen, H. (2017). Predictive processing increases intelligibility of acoustically distorted speech: behavioral and neural correlates. *Brain Behav.* 7:e00789. doi: 10.1002/brb3.789

Hasselmo, M. E., and Stern, C. E. (2014). Theta rhythm and the encoding and retrieval of space and time. *Neuroimage* 85(Pt 2), 656–666. doi: 10.1016/j.neuroimage.2013.06.022

Helmholtz, H. (1867). *Handbuch der Physiologishen Optik*. Leipzig: Leopold Voss.

Hobson, J. A., Hong, C. C., and Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Front. Psychol.* 5:1133. doi: 10.3389/fpsyg.2014.01133

Hohwy, J. (2013). *The Predictive Mind*. New York, NY: Oxford University Press.

Holmes, J., and Nolte, T. (2019). "Surprise" and the bayesian brain: implications for psychotherapy theory and practice. *Front. Psychol.* 10:592. doi: 10.3389/fpsyg.2019.00592

Iacoviello, B. M., Wu, G., Abend, R., Murrough, J. W., Feder, A., Fruchter, E., et al. (2014). Attention bias variability and symptoms of posttraumatic stress disorder. *J. Trauma Stress* 27, 232–239. doi: 10.1002/jts.21899

Jeffries, F. W., and Davis, P. (2013). What is the role of eye movements in eye movement desensitization and reprocessing (EMDR) for post-traumatic stress disorder (PTSD)? a review. *Behav. Cogn. Psychother.* 41, 290–300. doi: 10.1017/s1352465812000793

Johansson, R., Holsanova, J., Johansson, M., Dewhurst, R., and Holmqvist, K. (2012). Eye movements play an active role when visuospatial information is recalled from memory. *J. Vis.* 12, 1256–1256. doi: 10.1167/12.9.1256

Johansson, R., and Johansson, M. (2014). Look here, eye movements play a functional role in memory retrieval. *Psychol. Sci.* 25, 236–242. doi: 10.1177/0956797613498260

Jovanovic, T., Kazama, A., Bachevalier, J., and Davis, M. (2012). Impaired safety signal learning may be a biomarker of PTSD. *Neuropharmacology* 62, 695–704. doi: 10.1016/j.neuropharm.2011.02.023

Juan, C. H., Muggleton, N. G., Tzeng, O. J. L., Hung, D. L., Cowey, A., and Walsh, V. (2008). Segregation of visual selection and saccades in human frontal eye fields. *Cereb. Cortex* 18, 2410–2415. doi: 10.1093/cercor/bhn001

Jutras, M. J., Fries, P., and Buffalo, E. A. (2013). Oscillatory activity in the monkey hippocampus during visual exploration and memory formation. *Proc. Natl. Acad. Sci. U.S.A.* 110, 13144–13149. doi: 10.1073/pnas.1302351110

Kant, I. (1781). *Kritik der Reinen VERNUNFT (Critique of Pure Reason)*. Hamburg: Felix Meiner.

Kindt, M., Soeter, M., and Sevenster, D. (2014). Disrupting reconsolidation of fear memory in humans by a noradrenergic beta-blocker. *J. Vis. Exp.* 94:e52151. doi: 10.3791/52151

Kitamura, T., Ogawa, S. K., Roy, D. S., Okuyama, T., Morrissey, M. D., Smith, L. M., et al. (2017). Engrams and circuits crucial for systems consolidation of a memory. *Science* 356, 73–78. doi: 10.1126/science.aam6808

Kroes, M. C., and Fernandez, G. (2012). Dynamic neural systems enable adaptive, flexible memories. *Neurosci. Biobehav. Rev.* 36, 1646–1666. doi: 10.1016/j.neubiorev.2012.02.014

Kroes, M. C., Schiller, D., LeDoux, J. E., and Phelps, E. A. (2016). Translational approaches targeting reconsolidation. *Curr. Top. Behav. Neurosci.* 28, 197–230. doi: 10.1007/7854_2015_5008

Laeng, B., Bloem, I. M., D'Ascenzo, S., and Tommasi, L. (2014). Scrutinizing visual images: the role of gaze in mental imagery and memory. *Cognition* 131, 263–283. doi: 10.1016/j.cognition.2014.01.003

Landin-Romero, R., Moreno-Alcazar, A., Pagani, M., and Amann, B. L. (2018). How does eye movement desensitization and reprocessing therapy work? a systematic review on suggested mechanisms of action. *Front. Psychol.* 9:1395. doi: 10.3389/fpsyg.2018.01395

Lee, J. L. (2009). Reconsolidation: maintaining memory relevance. *Trends Neurosci.* 32, 413–420. doi: 10.1016/j.tins.2009.05.002

Linson, A., Parr, T., and Friston, K. (2019). Active inference, stressors and psychological trauma: a neuroetholigical model of (mal)adaptive explore-exploit dynamics in ecological context. *bioRxiv* doi: 10.1101/695445

Liu, Z. X., Shen, K., Olsen, R. K., and Ryan, J. D. (2016). Visual sampling predicts hippocampal activity. *J. Neurosci.* 37, 599–609. doi: 10.1523/jneurosci.2610-16.2016

McDermott, T. J., Badura-Brack, A. S., Becker, K. M., Ryan, T. J., Bar-Haim, Y., Pine, D. S., et al. (2016). Attention training improves aberrant neural dynamics during working memory processing in veterans with PTSD. *Cogn. Affect. Behav. Neurosci.* 16, 1140–1149. doi: 10.3758/s13415-016-0459-7

Meister, M. L., and Buffalo, E. A. (2016). Getting directions from the hippocampus: the neural connection between looking and memory. *Neurobiol. Learn. Mem.* 134 (Pt A), 135–144. doi: 10.1016/j.nlm.2015.12.004

Micic, D., Ehrlichman, H., and Chen, R. (2010). Why do we move our eyes while trying to remember? The relationship between non-visual gaze patterns and memory. *Brain Cogn.* 74, 210–224. doi: 10.1016/j.bandc.2010.07.014

Mirza, M. B., Adams, R. A., Mathys, C., and Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLoS One* 13:e0190429. doi: 10.1371/journal.pone.0190429

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Front. Comput. Neurosci.* 10:56. doi: 10.3389/fncom.2016.00056

Monfils, M.-H., Cowansage, K. K., Klann, E., and LeDoux, J. E. (2009). Extinction-reconsolidation boundaries: key to persistent attenuation of fear memories. *Science* 324, 951–955. doi: 10.1126/science.1167975

Naatanen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clin. Neurophysiol.* 118, 2544–2590. doi: 10.1016/j.clinph.2007.04.026

Nader, K., Schafe, G. E., and Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* 406, 722–726. doi: 10.1038/35021052

Naim, R., Abend, R., Wald, I., Eldar, S., Levi, O., Fruchter, E., et al. (2015). Threat-related attention bias variability and posttraumatic stress. *Am. J. Psychiatry* 172, 1242–1250. doi: 10.1176/appi.ajp.2015.14121579

Nobre, A. C., Gitelman, D. R., Dias, E. C., and Mesulam, M. M. (2000). Covert visual spatial orienting and saccades: overlapping neural systems. *Neuroimage* 11, 210–216. doi: 10.1006/nimg.2000.0539

Ohkawa, N., Saitoh, Y., Suzuki, A., Tsujimura, S., Murayama, E., Kosugi, S., et al. (2015). Artificial association of pre-stored information to generate a qualitatively new memory. *Cell Rep.* 11, 261–269. doi: 10.1016/j.celrep.2015.03.017

O'Keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* 34, 171–175. doi: 10.1016/0006-8993(71)90358-1

Parker, A., Buckley, S., and Dagnall, N. (2009). Reduced misinformation effects following saccadic bilateral eye movements. *Brain Cogn.* 69, 89–97. doi: 10.1016/j.bandc.2008.05.009

Parker, A., and Dagnall, N. (2010). Effects of handedness and saccadic bilateral eye movements on components of autobiographical recollection. *Brain Cogn.* 73, 93–101. doi: 10.1016/j.bandc.2010.03.005

Parker, A., Relph, S., and Dagnall, N. (2008). Effects of bilateral eye movements on the retrieval of item, associative, and contextual information. *Neuropsychology* 22, 136–145. doi: 10.1037/0894-4105.22.1.136

Parr, T., and Friston, K. J. (2017). The active construction of the visual world. *Neuropsychologia* 104, 92–101. doi: 10.1016/j.neuropsychologia.2017.08.003

Pedreira, M. E., Perez-Cuesta, L. M., and Maldonado, H. (2004). Mismatch between what is expected and what actually occurs triggers memory reconsolidation or extinction. *Learn. Mem.* 11, 579–585. doi: 10.1101/lm.76904

Preston, A. R., and Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Curr. Biol.* 23, R764–R773. doi: 10.1016/j.cub.2013.05.041

Przybyslawski, J., Roullet, P., and Sara, S. J. (1999). Attenuation of emotional and nonemotional memories after their reactivation: role of beta adrenergic receptors. *J. Neurosci.* 19, 6623–6628. doi: 10.1523/jneurosci.19-15-06623.1999

Raichle, M. E. (2010). Two views of brain function. *Trends Cogn. Sci.* 14, 180–190. doi: 10.1016/j.tics.2010.01.008

Rajkai, C., Lakatos, P., Chen, C. M., Pincze, Z., Karmos, G., and Schroeder, C. E. (2008). Transient cortical excitation at the onset of visual fixation. *Cereb. Cortex* 18, 200–209. doi: 10.1093/cercor/bhm046

Reinhart, R. M. G., and Nguyen, J. A. (2019). Working memory revived in older adults by synchronizing rhythmic brain circuits. *Nat. Neurosci.* 22, 820–827. doi: 10.1038/s41593-019-0371-x

Richter, F. R., Chanales, A. J. H., and Kuhl, B. A. (2015). Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *Neuroimage* 124(Pt A), 323–335. doi: 10.1016/j.neuroimage.2015.08.051

Rodenburg, R., Benjamin, A., de Roos, C., Meijer, A. M., and Stams, G. J. (2009). Efficacy of EMDR in children: a meta-analysis. *Clin. Psychol. Rev.* 29, 599–606. doi: 10.1016/j.cpr.2009.06.008

Sack, M., Lempa, W., Steinmetz, A., Lamprecht, F., and Hofmann, A. (2008). Alterations in autonomic tone during trauma exposure using eye movement desensitization and reprocessing (EMDR)–results of a preliminary investigation. *J. Anxiety Disord.* 22, 1264–1271. doi: 10.1016/j.janxdis.2008.01.007

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Schacter, D. L., and Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 362, 773–786. doi: 10.1098/rstb.2007.2087

Schiller, D., Monfils, M. H., Raio, C. M., Johnson, D. C., Ledoux, J. E., and Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* 463, 49–53. doi: 10.1038/nature08637

Scholz, A., Mehlhorn, K., and Krems, J. F. (2016). Listen up, eye movements play a role in verbal memory retrieval. *Psychol. Res.* 80, 149–158. doi: 10.1007/s00426-014-0639-4

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., and Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. *Curr. Opin Neurobiol.* 20, 172–176. doi: 10.1016/j.conb.2010.02.010

Schubert, S. J., Lee, C. W., and Drummond, P. D. (2011). The efficacy and psychophysiological correlates of dual-attention tasks in eye movement desensitization and reprocessing (EMDR). *J. Anxiety Disord.* 25, 1–11. doi: 10.1016/j.janxdis.2010.06.024

Scott, J. C., Matt, G. E., Wrocklage, K. M., Crnich, C., Jordan, J., Southwick, S. M., et al. (2015). A quantitative meta-analysis of neurocognitive functioning in posttraumatic stress disorder. *Psychol. Bull.* 141, 105–140. doi: 10.1037/a0038039

Scoville, W. B., and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* 20, 11–21. doi: 10.1136/jnnp.20.1.11

Sengupta, B., Stemmler, M. B., and Friston, K. J. (2013). Information and efficiency in the nervous system-a synthesis. *PLoS Comput. Biol.* 9:e1003157. doi: 10.1371/journal.pcbi.1003157

Sevenster, D., Beckers, T., and Kindt, M. (2014). Prediction error demarcates the transition from retrieval, to reconsolidation, to new learning. *Learn. Mem.* 21, 580–584. doi: 10.1101/lm.035493.114

Shapiro, F. (2001). *Eye Movement Desensitization and Reprocessing (EMDR): Basic Principles, Protocols, and Procedures*, 2nd Edn. New York, NY: Guilford Press.

Shapiro, F. (2018). *Eye Movement Desensitization and Reprocessing (EDMR) Therapy : Basic Principles, Protocols, and Procedures*, 3rd Edn. New York, NY: The Guilford Press.

Shapiro, F., and Laliotis, D. (2011). EMDR and the adaptive information processing model: integrative treatment and case conceptualization. *Clin. Soc. Work J.* 39, 191–200. doi: 10.1007/s10615-010-0300-7

Shen, K., Bezgin, G., Selvam, R., McIntosh, A. R., and Ryan, J. D. (2016). An anatomical interface between memory and oculomotor systems. *J. Cogn. Neurosci.* 28, 1772–1783. doi: 10.1162/jocn_a_01007

Siegle, J. H., and Wilson, M. A. (2014). Enhancement of encoding and retrieval functions through theta phase-specific manipulation of hippocampus. *eLife* 3:e03061. doi: 10.7554/eLife.03061

Sijbrandij, M., Engelhard, I. M., Lommen, M. J., Leer, A., and Baas, J. M. (2013). Impaired fear inhibition learning predicts the persistence of symptoms of posttraumatic stress disorder (PTSD). *J. Psychiatr. Res.* 47, 1991–1997. doi: 10.1016/j.jpsychires.2013.09.008

Sipos, M. L., Bar-Haim, Y., Abend, R., Adler, A. B., and Bliese, P. D. (2014). Postdeployment threat-related attention bias interacts with combat exposure to account for PTSD and anxiety symptoms in soldiers. *Depress. Anxiety* 31, 124–129. doi: 10.1002/da.22157

Smout, C. A., Tang, M. F., Garrido, M. I., and Mattingley, J. B. (2019). Attention promotes the neural encoding of prediction errors. *PLoS Biol.* 17:e2006812. doi: 10.1371/journal.pbio.2006812

Solomon, R., and Shapiro, F. (2008). EMDR and the adaptive information processing model. *J. EMDR Pract. Res.* 2, 315–325.

Stefanics, G., Heinzle, J., Horvath, A. A., and Stephan, K. E. (2018). Visual mismatch and predictive coding: a computational single-trial ERP study. *J. Neurosci.* 38, 4020–4030. doi: 10.1523/jneurosci.3365-17.2018

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiol. Behav.* 106, 5–15. doi: 10.1016/j.physbeh.2011.06.004

Stickgold, R., and Walker, M. P. (2007). Sleep-dependent memory consolidation and reconsolidation. *Sleep Med.* 8, 331–343. doi: 10.1016/j.sleep.2007.03.011

Strelnikov, K. (2007). Can mismatch negativity be linked to synaptic processes? A glutamatergic approach to deviance detection. *Brain Cogn.* 65, 244–251. doi: 10.1016/j.bandc.2007.04.002

Strelnikov, K. (2010). Neuroimaging and neuroenergetics: brain activations as information-driven reorganization of energy flows. *Brain Cogn.* 72, 449–456. doi: 10.1016/j.bandc.2009.12.008

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Teyler, T. J., and Rudy, J. W. (2007). The hippocampal indexing theory and episodic memory: updating the index. *Hippocampus* 17, 1158–1169. doi: 10.1002/hipo.20350

Treanor, M., Brown, L. A., Rissman, J., and Craske, M. G. (2017). Can memories of traumatic experiences or addiction be erased or modified? A critical review of research on the disruption of memory reconsolidation and its applications. *Perspect. Psychol. Sci.* 12, 290–305. doi: 10.1177/1745691616664725

Tulving, E., and Schacter, D. L. (1990). Priming and human memory systems. *Science* 247, 301–306. doi: 10.1126/science.2296719

Ullman, S. (2019). Using neuroscience to develop artificial intelligence. *Science* 363:692. doi: 10.1126/science.aau6595

Ullman, S., Harari, D., and Dorfman, N. (2012). From simple innate biases to complex visual concepts. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18215–18220. doi: 10.1073/pnas.1207690109

Vernet, M., Quentin, R., Chanes, L., Mitsumasu, A., and Valero-Cabre, A. (2014). Frontal eye field, where art thou? Anatomy, function, and non-invasive manipulation of frontal regions involved in eye movements and associated cognitive operations. *Front. Integr. Neurosci.* 8:66. doi: 10.3389/fnint.2014.00066

Visser, R. M., Lau-Zhu, A., Henson, R. N., and Holmes, E. A. (2018). Multiple memory systems, multiple time points: how science can inform treatment to control the expression of unwanted emotional memories. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 373:20170209. doi: 10.1098/rstb.2017.0209

Vossel, S., Mathys, C., Daunizeau, J., Bauer, M., Driver, J., Friston, K. J., et al. (2014). Spatial attention, precision, and Bayesian inference: a study of saccadic response speed. *Cereb. Cortex* 24, 1436–1450. doi: 10.1093/cercor/bhs418

Walker, M. P., and Stickgold, R. (2016). Understanding the boundary conditions of memory reconsolidation. *Proc. Natl. Acad. Sci. U.S.A.* 113, E3991–E3992.

Watts, B. V., Schnurr, P. P., Mayo, L., Young-Xu, Y., Weeks, W. B., and Friedman, M. J. (2013). Meta-analysis of the efficacy of treatments for posttraumatic stress disorder. *J. Clin. Psychiatry* 74, e541–e550. doi: 10.4088/JCP.12r08225

Wilkinson, S., Dodgson, G., and Meares, K. (2017). Predictive processing and the varieties of psychological trauma. *Front. Psychol.* 8:1840. doi: 10.3389/fpsyg.2017.01840

Wurtz, R. H. (2008). Neuronal mechanisms of visual stability. *Vision Res.* 48, 2070–2089. doi: 10.1016/j.visres.2008.03.021

Yokose, J., Okubo-Suzuki, R., Nomoto, M., Ohkawa, N., Nishizono, H., Suzuki, A., et al. (2017). Overlapping memory trace indispensable for linking, but not recalling, individual memories. *Science* 355, 398–403. doi: 10.1126/science.aal2690

Yufik, Y. M., and Friston, K. (2016). Life and understanding: the origins of "Understanding" in self-organizing nervous systems. *Front. Syst. Neurosci.* 10:98. doi: 10.3389/fnsys.2016.00098

Conflict of Interest: 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 Psychology

HYPOTHESIS AND THEORY
published: 21 November 2019
doi: 10.3389/fpsyg.2019.02599

The Gravity of Objects: How Affectively Organized Generative Models Influence Perception and Social Behavior

*Patrick Connolly**

Counselling and Psychology Department, Hong Kong Shue Yan University, North Point, Hong Kong

Friston's (2010) free energy principle (FEP) offers an opportunity to rethink what is meant by the psychoanalytic concept of an object or discrete mental representation (Ogden, 1992). The significance of such objects in psychoanalysis is that they may be superimposed on current experience so that perceptions are partly composed of projected fantasy and partly of more realistic perception. From a free energy perspective, the psychoanalytic (person) object may be understood as a bounded set of prior beliefs about a "platonic" sort of person that provides a free energy minimizing, evidence maximizing, hypothesis to explain inference about – or dyadic interactions with – another. The degree to which realistic perception supervenes – relative to a platonic person object – will depend upon the precision assigned to the sensory evidence (concerning the person) relative to the prior beliefs about a platonic form. This provides a basis for not only explaining projection and transference phenomena but also conceptualizing a central assumption within the object relations psychoanalysis. As an example, the paper examines the Kleinian theory of split good or bad part objects as affectively organized generative models (or platonic part-object models) formed in early infancy. This also provides a basis for building on work by Kernberg (1984, 1996) by conceptualizing the role of the part object(s) in a continuum of reality testing, from mild errors in perception that are relatively easily corrected, through borderline affective instability and frequent shifts between part-object experience, to psychotic failures of reality testing, where Friston et al. (2016) proposed that aberrant precisions bias perception to high precision false beliefs (here cast as platonic part objects), such as stable perceptions of others (and possibly oneself) as persecutory agents of some sort. The paper demonstrates the value that the history of clinical insights into psychoanalysis (including object relations) and a system-based approach to the brain (including the free energy principle) can have for one another. This is offered as a demonstration of the potential value of an "Integrative Clinical Systems Psychology" proposed by Tretter and Löffler-Stastka (2018), which has the potential to integrate the major theoretical frameworks in the field today.

Keywords: object relations, free energy principle, integrative clinical systems psychology, systems theory, social perception, psychoanalysis, development

OPEN ACCESS

Edited by:
Jim Hopkins,
University College London,
United Kingdom

Reviewed by:
Daniela Flores Mosri,
Universidad Intercontinental, Mexico
Karl Friston,
University College London,
United Kingdom

***Correspondence:**
Patrick Connolly
patrickconnolly@live.com

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *30 July 2019*
Accepted: *01 November 2019*
Published: *21 November 2019*

Citation:
Connolly P (2019) The Gravity of Objects: How Affectively Organized Generative Models Influence Perception and Social Behavior. Front. Psychol. 10:2599. doi: 10.3389/fpsyg.2019.02599

INTRODUCTION

A broad field of study in psychoanalysis focuses on objects, referring to a "mental" object and was described by Ogden (1992) as a discrete mental representation. In theory, objects could refer to mental representations of anything in our lived experience, but the field is primarily concerned with our mental representations of people, whether there are people around us, people who have been important to us in our past, or even ideal or prototypical people who exist only in our imagination.

There are a couple of key reasons that our mental representations of people are worth studying. One reason is that our social behavior may be very influenced by those representations. The way we see people (either people in general, or specific types of people like authority figures, or specific individuals) may strongly influence our behavior toward them. Of course, many factors play a role in any given (social) behavior, including our aims toward others, strategic goals we are trying to achieve, our emotions, and a host of contextual factors. But, a key assumption of object relations psychology suggests that those representations do play quite a foundational role in our behavior, our aims toward people to some extent, and the emotions we feel about them.

Another key reason for interest in mental representations of people is because such representations may be quite different from the reality of the people that they are supposed to represent. Object relations psychologists suggest that a lot of our apparently irrational, self-destructive, or maladaptive behaviors – particularly in social contexts – may make perfect sense when we can see that people are acting consistently with their (often inaccurate) inner representations of people.

Following Kernberg (1965, 1987), a person's reaction to another person's behavior is potentially the result of two different components (Kernberg's idea was primarily applied to counter transference, or how therapists perceive clients, but logically applies to any interaction). The first of these is a "realistic" response to the person's behavior toward us, including their attitude and emotions toward us. In other words, someone's behavior might make me feel angry because it typically produces anger in others too, for example, where the other person insults me. We might think of my angry response toward them as a compatible or "rational" response to the other person's behavior in such a case.

However, my response may also be driven primarily by "fantasy" or the influence of my own mental representations rather than by a realistic perception of the stimulus[1]. In other words, I may have a response to the other person's behavior that few other people would have, or my response would be much more (or less) intense than other people's, beyond a level of what could be described as culturally or statistically normal. An example here would be an angry response to another person whose behavior would not ordinarily cause others to behave angrily: another person makes an inoffensive joke that does not really make any reference to me (or my social identity) in any way.

Object relations psychoanalysts (such as Klein, 1946; Kernberg, 1965; Ogden, 1992) tend to suggest that this kind of error in perception is ubiquitous. In other words, at all times, our perception of people may reflect a combination of both the realistic perceiving and some amount of "representation-driven" perceiving. All that vary are the relative *extent* (or ratio) of the realistic versus representation-driven perceiving. For most people, the relative influence of realistic versus representational perceiving varies from one situation to another. However, there are a number of conditions in which we may think that some people regularly have a much stronger influence of representation-driven perception. One instance may be personality disorders, such as borderline personality disorder or paranoid personality disorder, where perceptions of other people are mostly negative. An even more extreme example may be schizophrenia, where a person may remain entirely convinced of seriously hostile intentions of almost everyone around them, despite being exposed to a large amount of information that might appear to contradict such a perception. In each of these cases, the perception driven by the internal mental representation appears less responsive to the information available.

However, regardless of whether we are focusing on typical or atypical social perception, an important question raised by the above account is how such a "quantitative" description (in other words "more" or "less" based on available information) is formalized, and in a neuronally plausible way, which also fits the phenomenology we observe. This is computational in the sense that it describes an outcome (the perception of a person) as the result of two processes (realistic versus representation-driven perception) that appear to operate in opposition to one another, such that the result reflects some relative ratio of both. A computational account would require that the processes be specifically defined as quantifiable terms in an equation that precisely specifies the relation between them.

Friston's (2010) free energy principle (FEP) as a regulatory principle of biological organisms, including brain processes, offers precisely such a computational expression of the relative influence of informational inputs to the brain and how they are acted upon by the existing mental representations of persons encoded in its neural networks. This paper will outline a free energy principle account of person perception, showing how it is computationally efficient for the brain to encode a prototypical model of what a person is. This prototypical model then exerts a theoretically quantifiable influence on conscious perception. The focus then shifts toward showing how object relations theory in psychoanalysis can make use of this FEP account to demonstrate the unique contribution it can bring to formalizing social perception. The paper uses the example of good and bad part objects found by Klein's (1946) foundational work in object relations and suggests that these could be understood as distinct affectively organized generative models that play a role in social perception and that come to the fore in emotionally intense experiences. Next, this account of affectively organized part-object models is applied to both borderline personality

[1]Strictly, this is incorrect – all perception must be in terms of mental representations, and there is no direct perception of reality. Rather, what is implied are poles of a dimension, that is apparently more reality driven versus representation driven.

disorder and schizophrenia to show how this reformulated object relations approach might provide additional explanatory power to current computational FEP-based approaches to these psychiatric conditions. Finally, it is suggested that the formulation in the paper attempts to demonstrate the potential value of Tretter and Löffler-Stastka's (2018) call for an "Integrative Clinical Systems Psychology" that has the potential to integrate existing major theories of psychology with system-based approaches. We begin with laying out a formal free-energy principle account of person perception.

A PLATONIC PERSON MODEL APPROACH TO FREE-ENERGY PRINCIPLE-BASED SOCIAL PERCEPTION

A FEP account of information processing proposes that the physical structure of the brain constitutes a generative model of its environment that actively infers the causes of its sensory inputs and specifies a prior prediction of inputs. The organism (and the brain) acts according to a regulatory principle, which minimizes the differences (more correctly the Kullback-Leibler divergence, or "free energy") between the prior prediction of the generative model and the posterior likelihood of the inputs[2]. This minimization is achieved either through the Bayesian updating of the generative model or by an action, which alters the inputs in line with the predictions of the generative model. In this way, the free energy in the perceiving system drives both behavior and learning (i.e., belief updating).

A more complete account of free energy minimization rests on mathematically decomposing free energy in a number of ways. First, free energy can be decomposed into expected energy minus entropy. This means that minimizing free energy conforms to (Jaynes) principle of maximum entropy (Banavar et al., 2010). A more intuitive decomposition splits free energy into complexity minus accuracy. This means that minimizing free energy is equivalent to providing an accurate explanation for the sensorium in a minimally complex way – in accordance with Occam's principle (Maisto et al., 2015). Finally, free energy can be expressed as an evidence lower bound minus the log evidence for the generative model. This decomposition means that minimizing free energy reduces the bound (to ensure that model evidence is maximized). This is sometimes referred to as self-evidencing (Hohwy, 2016). These decompositions are mathematically equivalent; however, the decomposition into accuracy and complexity will figure prominently in the present discussion and is explained next.

The learning (belief updating) process described above always moves in the direction of greater accuracy of the model while minimizing complexity; namely, an oversensitivity to typical changes in inputs. In other words, my generative model of the world grows in accuracy with experience, but once it becomes too accurate (in a sense over-fitted to the data), it is becomes sub-optimal in that relatively small shifts in the state of the environment can now generate larger amounts of free energy (or prediction error). Therefore, it is computationally efficient for our generative models to be abstracted from our sensory experienced to some extent, so that they maximize accuracy while minimizing oversensitivity in typical changes in inputs (Friston, 2010; Hobson et al., 2014). Friston (2017, personal communication) suggests that this would be enough to explain the emergence of a mental representation (generative model) of a "platonic" person[3] encoded within the structure of the body and nervous system:

> "… a prior belief about a … 'platonic' sort of person provides a free energy minimising, evidence maximising, hypothesis to explain inference about – or dyadic interactions with – another. In other words, having a particular hypothesis or platonic person in mind allows you to immediately explain prosocial cues in an accurate and parsimonious fashion. The parsimony afforded by the object minimises complexity and thereby free energy."

This computationally efficient set of expectations cohering around an abstracted "platonic" model of a person informs our expectations regarding what people do and how they think and behave. In terms of how this generative model of a platonic person is encoded in the brain, this must of necessity refer to distributed networks of neural relationships within a multi-level hierarchy of organization, consisting of faster sensory-level priors and increasingly slower, more abstract integrative priors extending through the highest levels of cortical organization[4].

A FREE ENERGY FORMULATION OF REALITY VERSUS REPRESENTATION-DRIVEN PERCEPTION

The FE formulation of the platonic object as described above now allows for a formal statement of the relationship of reality versus representation-driven perception. Friston (2017, personal communication) describes how this distinction might be formulated in a free energy perspective:

[2]This formulation and the equation specifying the free energy principle can be found by Friston (2010).

[3]This refers to Plato's notion of an ideal form of a thing. For example, one might consider all the persons you have encountered in reality as reflections of a prototypical form of a person, which we might call a "platonic" person. The author thanks Dr Jeremy Holmes, who commented on an earlier version of this paper, and suggested the terms "schematic" or "stereotypical" as alternatives to "platonic" here.

[4]Diaconescu et al. (2017, submitted) presented evidence for a temporal sequence of activation of a proposed hierarchy of levels of cortical function for a social perception task that required different levels of processing of social information. The findings suggested that all areas typically associated with "theory of mind" tasks were activated during the sequence, such as the middle cingulate gyrus, medial prefrontal cortex, and temporo-parietal junction.

> "The degree to which realistic perception, relative to a platonic person object supervenes will depend upon the precision (usually cast as attention) assigned to the sensory evidence (concerning the person) and the prior beliefs about a platonic form. In other words, your posterior beliefs – following an encounter with another – will be a mixture of the object prior and the likelihood that the object is behaving in a way consistent with that hypothesis – or an alternative object."

In other words, this quantitative relation between reality and representation-driven perception is described in terms of probability. The posterior belief, which reflects our experience of a person, is partly determined by the information we receive, and the relative extent to which it matches (or does not) our existing prior. However, it is not just the information itself that determines this relative probability; as indicated above, it is also the precision afforded to the prior prediction. What precision means here, is the confidence assigned to the higher-level predictions of our generative model – if its high relative to the precision afforded to the sensory evidence, discrepancies with the sensory evidence will be attenuated to some degree, and vice versa.

This description can be thought of intuitively in terms of a formal similarity with gravity. In other words, our prior predictions exert a kind of influence on the perceived information, where it is stronger we tend to perceive our "representation" (object prior) and where it is weaker, we may perceive more of the reality, in so far as it is different from our inner representation. Friston (2017, personal communication) suggests:

> "The 'gravitational pull' of the object is, exactly the relative precision afforded to the prior hypothesis of the object, relative to the sensory evidence. In fact, mathematically, the equations that govern the posterior expectation have exactly the form used in Newtonian mechanics and gravitation."

The implication of this perspective is that our perception of people will always tend to some extent toward our platonic model of the person.

If we assume a hierarchical recursive development of the platonic model of the person (Connolly and van Deventer, 2017), we understand each new level of organization of this model to be constrained to some extent by what came before. In other words, our earliest experiences of people in a sense lay the foundation for the future development of the model.

The account of errors in perception offered thus far is not formally similar only to psychoanalytic theory. It is an established idea in cognitive theory that our social perception is shaped by schematic representations of people, which may lead to inaccurate information processing and maladaptive behavior. It may well be that an integrative clinical systems theory (Tretter and Löffler-Stastka, 2018), which is potentially able to incorporate a system paradigm such as Dr. Friston's work, may well be able to integrate these different theoretical perspectives, and more is said on this toward the end of the paper. However, the value of the field of psychoanalysis to the future growth of an integrative clinical systems paradigm lies in its sizeable literature of clinical insights that offers the possibility of better models, or descriptions, in this case of our person perception. In this regard, one of the most immediate contributions that psychoanalytic theory can make to the current FEP formulation of person perception is the observation that most minds contain more than one such platonic person object.

MULTIPLE OBJECTS

While the body of object relations theory may describe many different taxonomies of person objects, for the sake of clarity, this article will focus in detail on just one in order to unpack how a formal description might work and to examine its potential implications. In Klein's (1946) seminal text "Notes on some schizoid mechanisms," she proposed that in the earliest months of an infant's life, the child did not yet have an integrated (platonic) person object with which to perceive people as unitary, complex objects. Rather, we perceived only "part" objects, which are incomplete and fragmented representations of people, such that we might perceive a particular person at times as one part object, and at other times, a different one. She focused on part objects defined by good and bad experiences. Primarily, she focused on the child's experience of the mother's breast, defining the "good breast" as a founded upon a good experience of the breast as satisfying and pleasurable. By contrast, the "bad" breast was founded upon unsatisfying, frustrating, or withholding experience, into which the child projected their hostile emotions, borne of those experiences. She then described how, beginning from roughly 6 months of age, the child began to integrate those part objects into a whole object representation and perceives the mother as a whole person (Klein, 1946).

While Klein focused on the breast as a part object, it is broadly understood that these representations, fragmented as they are, are nonetheless part representations of different aspects of persons, though they are not yet perceived as belonging to the same object, the whole person. In other words, my experience of the bad mother part object is not yet integrated with my experience of the good mother part object. These must reflect different prototypes of platonic persons that are encoded as distinct generative models at this early stage of development.

The idea that an organism can encode distinct generative models for different "others" has been described by Isomura et al. (2018, unpublished). In their paper, they describe a theoretically and neurobiologically plausible model whereby a bird may fit sensory inputs from other birds under distinct

generative models. In this way, they may "know who they are communicating with," which allows for appropriate inferences within complex social environments.

While the formulation of Isomura et al. (2018, unpublished) might support the idea that we have different generative models for different "people," it should be made clear that Klein's theory of the good and bad object is not referring to different individuals (though it may initially), so much as it refers to different *types* of person or rather different platonic persons. In this case, these distinct platonic objects are founded upon different emotional experiences (pleasure and satisfaction versus dissatisfaction and frustration).

An indication of how such models may build upon a base of emotions, Panksepp's (1998) work on affective systems of mammalian brains is useful here. Panksepp described seven affective (command) systems common to mammalian brains, activation of which was associated with observable affective states and related behaviors. He presented these as in terms of core affective descriptions such as RAGE, LUST, or SEEKING and described the neural systems that appeared related to each of these. While Panksepp (1998, 2010) has offered a lot of evidence for his claims, there have nonetheless been criticisms, including some regarding the complex expression of human affect (Barrett et al., 2007). However, Panksepp (1998) did express the hope that 1 day the role of the affective systems would be understood within a broader system-based understanding:

> "The basic emotional systems may act as 'strange attractors' within widespread neural networks that exert a certain type of 'neurogravitational force' on many ongoing activities of the brain" (p. 3).

Here we see that affective systems may have their own "gravitational force" that entrains[5] the activities of the brain within their ambit, that is, at first (if Klein is correct), a stronger attractor than a platonic person object (which does not yet fully exist in the infant's brain). Fitting this description within a hierarchically recursive scheme of the nervous system, we could say that the seven affective command systems described by Panksepp have a tremendous influence on higher-order functioning of the brain, acting as constraints on superordinate levels of processing, such that our earliest experiences of people may be updating generative models that were originally distinguished by affect, just as Klein suggests.

THE EMERGENCE OF A DOMINANT PLATONIC PERSON

Around 6 months of age, Klein (1946) suggested that the child began to integrate these affectively defined good and bad objects within a "whole-object" representation in which the mother is perceived as a unitary person. She also proposed that there began a meaningful shift in the affective relationship with the whole-mother object that tended to move away from extremes of persecutory anxiety, rage, and intense idealization, toward a more ambivalent relationship characterized by guilt and reparation. This also marks the beginning of our capacity for more realistic object relating. We might say that the gravitational pull of the generative model that is organized by experiences of the whole object slowly begins to exceed that of the part objects that were organized by the affective command systems. However, this is better described in the sense of a recursive hierarchy (Connolly and van Deventer, 2017), where perception is originally entrained by the affective systems but later both perception and the affective systems come to be entrained by the increasingly stable object organization, which has emerged as a superordinate level of organization to that determined by the affective systems, though is still constrained by them.

This state of affairs is represented in the hierarchical matrix in **Figure 1**, which is adapted from one found by Tretter and Löffler-Stastka (2018), which displayed the core concept of object relations theory (Kernberg, 1976) as a matrix, though their figure included a description of the emergence of self from the environment, while the present figure focuses on describing the emergence of an object representation from a background of other experiences (including of the self), through a recursive development[6].

The view being given in this paper is that prior to the emergence of a dominant platonic person object, there are some number of distinct generative models (that predict sensory inputs), which are largely organized by the affective systems (depicted in **Figure 1** by the second layer from the bottom). The question here is how a dominant platonic person object comes to emerge.

Here, we can borrow terminology from systems theory by describing the emergence of distinct generative models for person perception as a progressive segmentation occurring within the broader generative model, where one of the parts

[5]The slaving principle in physics (Haken, 1983/2004) proposes that slower "macro" processes entrain faster microprocesses. This concept has been applied to levels of neural architecture (Badcock et al., 2019) and levels of organization of the mind in psychoanalysis (Connolly and van Deventer, 2017; Connolly, 2019). The dominant platonic person model, as a much more complex (and accurate) model involving more complex functional connectivity, must also be a slower, "bigger" process than part-models and can entrain them. The present description obviously parallels Kahneman's (2011) "fast and slow thinking" processes.

[6]This paper has not focused explicitly on the link between one's self-representation and our object-representations of others. Clearly this is of some importance as there may be some basis for suggesting that self- and other-representations overlap to some extent; for example, Singer et al. (2004) found that activations while apparently observing significant others receiving a pain stimulus overlap with activations when experiencing that same stimulus, though there may be distinct mechanisms for perceiving self and other (Lamm et al., 2016). Even potentially distinct self- and other-representations may influence one another in person perception. Moutoussis et al. (2014a,b) have offered an active inference model of person perception using a simplified Trust game task, which showed that self-representations in the form of preferences of the sort of person I am influence my goals in interaction as well as how I see the other. In the same way, I may infer what sort of person I am (or will be) from how the interaction progresses. This interaction between self- and other-representation requires further development than given in the present paper.

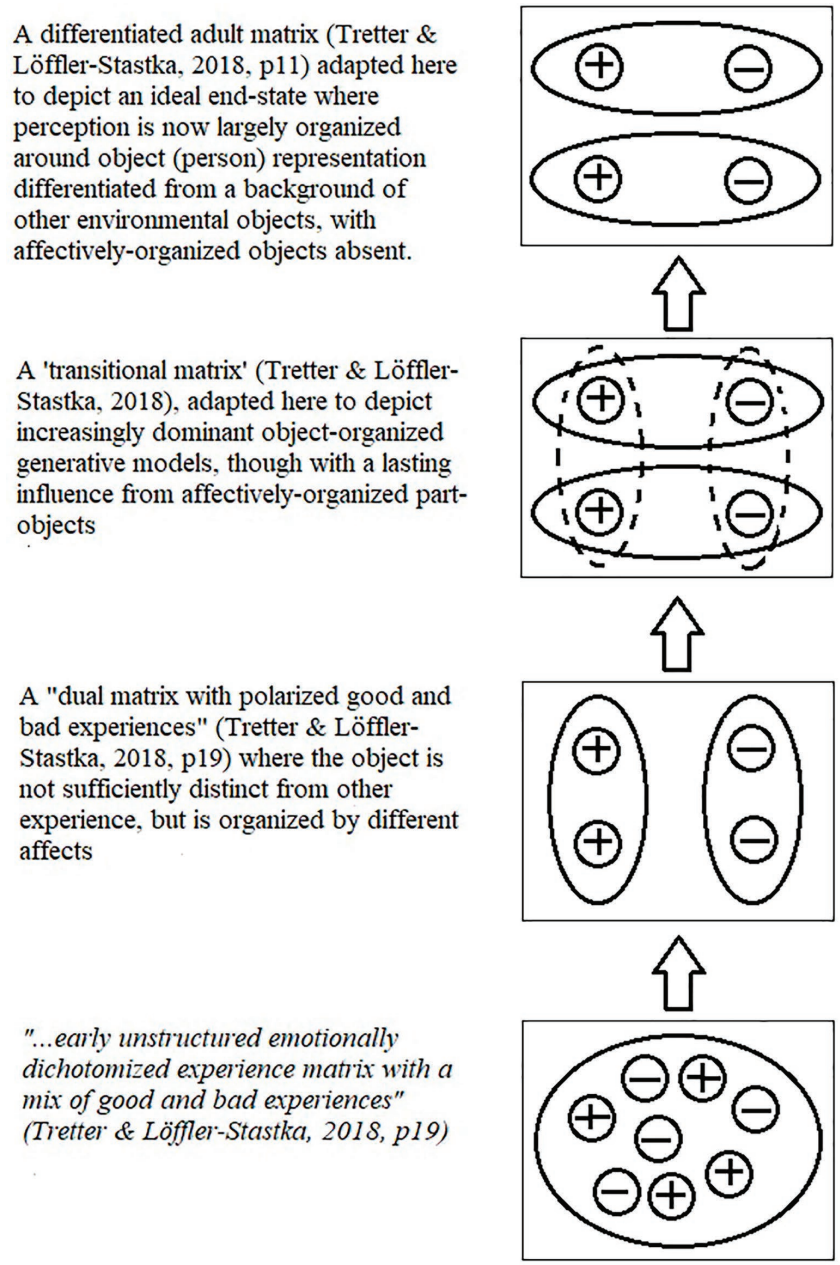

FIGURE 1 | The emergence of person object representation differentiated from other representations, from a lower level of affectively organized part objects (adapted from Tretter and Löffler-Stastka, 2018, p. 11, with permission from the copyright holder).

comes to lead, through a feedback loop with the sensory data, until it is dominant, a state of affairs depicted in **Figure 2**.

It is important to note that what is being proposed is that dominant platonic person representation has been built upon "part objects" that were previously affectively organized. It might be tempting to suggest that it is a model built upon one particular affective system that comes to dominate. For example, since we think that maturation involves an increasing ability to perceive people without much apparent emotion, we might suggest that models built upon the SEEKING system steadily come to dominate.

However, it is very unlikely that the dominant platonic person object is founded on the activity one particular affective system. Rather, given the phenomenology we observe in people, it must be a more complex mixture of the pre-existing generative models (the part objects).

From the formal perspective of minimizing free energy – or maximizing model evidence, the hierarchical assembly of parsimonious models of the (prosocial) world through our development can be considered in the light of Bayesian model selection or what is called "structure learning"

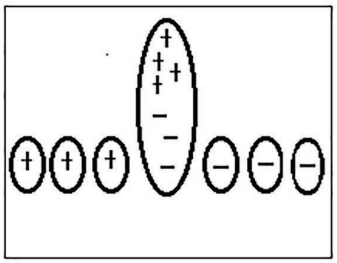

FIGURE 2 | The emergence of a dominant platonic person model (integrating positive and negative affects) as a leading part from a layer of part-object models organized by affect.

(Gershman and Niv, 2010; Tervo et al., 2016; Isomura and Friston, 2018). In other words, one can build more comprehensive (deep) generative models that have greater evidence (i.e., accuracy minus complexity) by adding layers or rearranging part objects into more complex (or deeper) wholes. These operations of hierarchical assembly, for example, "split" and "merge" operators, figure prominently in machine learning and statistics – and may provide a nice metaphor for the merging of part objects into more complex (or more dominant) objects or, as will be shown later, the "splitting off" phenomena seen in psychoanalysis.

This tendency toward hierarchical assembly of existing part objects may also imply that the emotions that we tend to experience often during early development tend to have a greater organizational influence (in the form of constraints) on the dominant platonic person model. For example, frequent experiences of PLAY affects in our early development are likely to influence the platonic person object toward perceiving people as fun, while frequent experiences of FEAR are likely to influence it toward perceiving people as dangerous. Each new experience characterized by these emotions increases the influence they have over our social perception. Over time, where the environment permits, these are likely to become stable, self-organizing perceptions of people[7]. This description also suggests how our dominant person model may form templates of different types of persons, partly constructed from different combinations of pre-existing part objects.

THE PERSISTENCE OF PRIMAL OBJECTS

The section above has offered a theoretical formalization of Klein's (1946) assertion that we come to perceive people as whole objects, so that (to some extent at least) we perceive that a person is the same person regardless of the emotions we have toward them. However, the capacity to "split" our perceptions of people around us into good and bad objects appears to be an ongoing phenomenon well into adult life, especially in particular situations, in which it is understood as a "splitting" defense. We may perceive a competitor in an intense rivalry as "all bad," or a new lover as "all good," or split the representation between people, for example, two teachers at school, one all good, and one all bad. These states supposedly reflect the persistence of the good and bad part objects as "latent" objects in the organization of the psyche that may nonetheless come to the fore in situations that activate them.

We might say that the dominant platonic person object gains profoundly greater influence over perception relative to the generative models of the affective (e.g., good and bad) part objects, and it is plausible that those part objects remain present as an influence in the nervous system provided that the connections that encode them experience reconsolidation at least occasionally.

The influence such "primal" generative models have on conscious perception[8] may vary from extreme to fairly subtle. At the extreme end are experiences of the most intense emotions, where we seem to experience almost nothing else but that emotion, with little higher thought process. An example is a man going through an acrimonious divorce process who encounters his ex-partner at a shopping center, accompanied by a new lover. He later describes himself as overwhelmed and rooted to the spot, at that moment feeling as if the universe had just shattered in some way, as everything else faded into the background and he only saw her laugh, her hand on her new partner's arm, and experienced her only as a terrible beautiful thing that was tearing his body apart from the inside. Shortly afterward, a more normal thought process resumes though he feels shaken and very distressed.

We can suggest that a part object has pulled perception entirely within its event horizon, and the dominant platonic

[7]My thanks to Dr. Pieter Grobbelaar, who suggested that at an early age, the self becomes constituted around either predominantly positive emotions or predominantly negative emotions, which tend to persist through the lifespan. This fits the view being expressed in this paper, in which affective experience is organizing experiences of both the self and the others. Given the adaptive value of positive affect in child development and adaptation, caregivers of very young children who wish to facilitate future positive affect in children should aim to maximize the child's experiences of pleasure and minimize negative emotional experiences in the earliest stages, as far as that is possible.

[8]The view of consciousness taken in this paper follows the description by Hobson et al. (2014) of the highest level of a hierarchy of organization, which is founded upon counterfactual simulation of future consequences of actions of sufficient depth (Friston, 2018).

generative model seems to have lost all influences on perception during this experience. At that moment, the man does not perceive his ex-partner as a "person" at all, but only as some surface sensory characteristics, distressing interoceptive sensations and an inarticulate sense of persecution.

The key point being made here is that there may already be a pre-existing generative (part-object) model, which represents the best prediction of the combination of sensory and interoceptive input at this present moment. The most typical hypothesis among object relations thinkers is usually that there are experiences of early childhood that were of similar emotional valence, which are activated by the contemporary experience. For example, feelings of abandonment (likely related to activation of PANIC/GRIEF affective functioning, as described by Panksepp, 2010) connected to several early experiences of the man's mother regularly going to work in the morning after only a few months of maternity leave, chatting with his father as she walked out the door, and other similar early experiences that organized around those emotions.

The reason for why ordinary person perception seems to be suspended may be partly due to the fact that the thought processes are state dependent to some extent, and most people have very limited experiences (and generative models) at such intense emotional valences[9]. These state-dependent thought processes, emotions, and body feelings may also compete with more ordinary person perception for activation of shared networks, such as described by Oosterwijk et al. (2012).

A further consideration relates to Freud's concept of conflict, where aspects of our psyche no longer undergo normal development due to their generating too much conflict during development and remaining repressed (Freud, 1912/1963, 1915/1963). This has been formally described by Hopkins (2016) and elaborated by Connolly (2018), as a situation where alternative plans of action generate similar high levels of expected free energy. Through development, the "loser" of this contest becomes progressively less able to determine high-level conscious experience. In this way, the child's distressing experiences related to feelings of abandonment (PANIC/GRIEF affects) when their mother left for work may likely lead to policies of action (such as rejecting the abandoner) that also generate high expected levels of free energy. To resolve the conflict, the superordinate levels of the person's generative model alter the precisions afforded prior beliefs about policies of action, where these prior beliefs are based on the expected free energy following a particular action. Should the distressing feelings related to PANIC/GRIEF lose the competition, they become less and less likely to be activated in the normal course of affairs, and the dominant platonic person model that emerges through further development is likely to encode this constraint. In this way, the part model becomes a "split-off" remnant that is not integrated into the dominant platonic person model and does not undergo the significant further updating that the dominant model does, though it may come into association with experiences of similar emotional valence (Freud, 1915/1963), which seems to be what happens to the man in the example[10].

The rubber hand illusion provides a nice metaphor for this sort of process from a free energy perspective[11]. In the rubber hand illusion, concomitant visual and tactile information is supplied *via* stroking a rubber hand, inducing the illusion that the hand is part of one's body. The most common explanation – for this illusory body ownership – is that the proprioceptive (position) sensory information that is attenuated (i.e., ignored) by reducing its precision (Paton et al., 2012; Seth, 2013; Zeller et al., 2015). This enables a low free energy explanation for the coherent visual and tactile information under the belief that "I only have one right hand." In short, the high level prior beliefs about the part objects that comprise my body can have a profound effect on the way in which evidence is accumulated for those beliefs, under active inference. In this way, the attenuated proprioceptive sensory information is akin to the split off remnant described in the paragraph above, in which it is no longer consciously experienced in the ordinary state of affairs. Rather, similar to the high-level beliefs about one's hand, beliefs about a dominant platonic person model come to have greater precision and begin to shift the accumulation of evidence in line with this prior.

This description of the early formation (and splitting off) of the generative part-object model might also offer a hypothesis to explain the dissociated or "de-realized" characteristics of the experience, where the surrounding reality, sense of self, and ordinary thoughts are somehow not perceived consciously. These earliest part-object models formed in early stages of development where functional connectivity is far less developed. For example, in research that later led to a Nobel prize, Hafting et al. (2005) reported the activity of grid cells that provided a sense of place throughout all experience. More recently, Tsao et al. (2018) have shown that cells in the lateral entorhinal cortex encode a perception of time in experience. While we are born with these structures, their successful integration with conscious perception is surely a developmental achievement. It seems possible that the seemingly "derealized" nature of these experiences may result because these part-object models formed before such complex integration has fully taken place. Of course, it may simply be explained rather by the intense emotional valence and demand for network resources meaning

[9]Eryilmaz et al. (2011) found changes in functional connectivity impacting on resting states following transient emotion.

[10]In a related way, horror movies depicting demonic characters who are only motivated by extreme sadistic empathy may well be so frightening to people because they activate such early part objects organized around emotions of persecutory anxiety or fear. Freud (1919/1955) offered an idea like this in "The Uncanny," which refers to "… that class of the frightening which leads back to what is known of old and long familiar." (p. 220). This description of affectively organized generative part-object models may also offer some new life for Jung's concept of *archetypal figures* in psychology. It would seem interesting to explore how figures such as the great mother, the child, the devil, and the trickster may be related to early experience organized by corresponding (mixtures of) activation of affective command systems. The unique imagery or thematic nature of our personal part objects is a result of our ongoing experiences (some of which are culturally shared) that have updated these models to a small extent.

[11]My thanks to reviewer, Dr. Karl Friston with constructive assistance with this and other points in the paper.

that the activity of these orienting systems is temporarily not integrated with conscious perception. However, the present formulation offers an alternative hypothesis, and, of course, both may simultaneously be true.

The above descriptions have referred to situations where primal part models influence conscious perception in an extreme sort of way. However, their influence may run on a continuum down to more subtle influences. This refers to situations where our dominant platonic person generative model is largely engaged in active inference of a social situation, but platonic part models still "drag" the perception in their direction to some extent.

As an example, we could refer to the same man as in the example above, though at an earlier point in his marriage, before the divorce. During dinner, his wife answers a phone call, says it is a work colleague, and steps outside and has a long laughing chat on the phone, leaving the husband to eat alone with their children. The man feels irritated by this, but thinks no more of it. However, he finds he is irritable with a number of his wife's behaviors for the rest of the evening, perceiving a lack of care or consideration in several behaviors. Only after some reflection does he realize that it began with the phone call.

In this instance, it may be that the platonic part model is activated, but unable to have the same dramatic influence on conscious perception. Instead, its influence on conscious perception can be thought of in terms of the binocular rivalry paradigm as presented by Hopkins (2012) where competing predictions about different stimuli presented to each eye (e.g., whether it is a house or a face) seem to dominate in cycles. Following that example, one can think of the dominant platonic person model as being in competition with the platonic part model to explain the current stimuli (in this case, the wife stepping out of dinner to chat with a colleague). The part model may not be dominant enough to come to define conscious perception as it did in the more extreme example above (or in the binocular rivalry example) but may still generate some lower level free energy within the psyche, which may not be adequately explained by the model dominating perception. This activation of the part model and its accompanying affect sets up a feedback loop where it becomes sustained over the rest of the evening, where the man's continuing experiences of his wife throughout the evening trigger inferences to explain the negative affect (inferences related to perceiving her behavior as "abandoning" him or not considering him), which reactivates the part model, and so on.

This feedback loop seems to explain how, once we dislike a person, we may often struggle to shift into liking them, particularly when we do not really know why we dislike them. The negative affect emerging from whatever negative part models activated by our experience of that person seems to result in an ongoing process of negative inferences about the person's behavior that feed back into the reactivation of the underlying part models (if they are involved in the dislike), even though we may never have a conscious perception of what we really feel about the person, and why. Having said this, we do nonetheless have experiences of being able to escape a more transient affectively influenced perception of a person, which is addressed next.

DOMINANT VERSUS PART MODELS, NOT COGNITION VERSUS EMOTION

We do have experiences where we have a strong emotional reaction to a person (and an attendant set of inferences), and then seem to have an insight or a new thought about it that seems to reinstate an apparently more objective perception. An example might be encountering a cashier in a supermarket who seems to be surly and unpleasant in handling our transaction and rolls his eyeballs when we drop an item by mistake. We may have a strong negative reaction and make an irrational inference that he dislikes us personally and is attacking us. Then, we suddenly think that it is likely that he has a bad mood (and perhaps often does) or treats almost everyone like that. We may even wonder about what troubles he may have in his life that makes him unhappy. This can seem to make our intense emotional response (and personalized inferences) evaporate fairly immediately.

It is this form of "reality-testing" process that is one of many observations that seem to support the idea that cognition is in competition with emotion to determine our perception, and the potential that higher cognitive process has to influence emotion-driven thought processes. However, it may be that "emotion versus cognition" is not the best distinction to draw in this situation. Rather, we can describe a competition between a highly developed dominant platonic person model and a more archaic, underdeveloped part-object model, to regulate free energy.

The more extreme activation of the part model in the earlier example of the cashier is distinct precisely because the more complex dominant model *fails* to entrain the activation of the part model (for a brief period at least) because this part model has never been integrated within the dominant model for the reasons described earlier in the paper. However, the much greater pull of the dominant model soon reasserts itself. It seems clear that through development, as our platonic person model (and the functional connectivity that underwrites it) ascends in complexity and ever more accurately "recognizes" these states, it entrains our affectively organized experiences more effectively as well and reduces their free energy. Besides this influence of general development of the dominant platonic model, the tendency toward reducing psychological conflict through altering the precisions regarding prior beliefs of expected free energy ascribed to policies of action (Hopkins, 2016; Connolly, 2018) related to part-object models means that they are increasingly avoided as the superordinate levels of the brain hierarchy that encode those precisions also develop. This may underlie a tendency toward reduction of the frequency and duration of intense part-object experiences in life, though it is a journey that may never entirely be complete.

However, in contrast to this tendency, a trend in research into psychopathology has focused on more severe deficits in functional connectivity underlying problems of active inference (rather than the more typical phenomena described above).

Disorders such as autism, schizophrenia, and personality disorders have been cast as problems of social inference. The next sections of this paper seek to apply the formulation that has been developed this far to briefly outline the potential it can have to contribute to our understanding to this approach to both borderline personality disorder (BPD) and schizophrenia. In doing so, the author attempts to place these phenomena on a continuum in terms of the relative stability of the dominant platonic person model in perception (or oppositely, the relative influence of part-object perception) and by implication, the level of effective functional connectivity that supports the dominant model. The subtle problems of reality testing described above refer to relatively higher dominant model, low part-object model perception (and relatively more normal connectivity), while BPD (examined next) is cast as more severe problems in maintaining the dominant model that entrains our perception, and schizophrenia representing the most severe problems in entraining part models in perception (and the most severe problems of connectivity)[12,13]. The distinction between BPD and schizophrenia is given here in terms of functional differences in terms of the level and stability of part-object object perception, and by implication, the level of functional connectivity, though these disorders may have discrete patterns of neurophysiological presentation and aberrant connectivity as well.

This presentation of disturbances in object perception on a continuum of levels of dominant versus part-object perception is both consilient with and inspired by a formulation by Kernberg (1984, 1996, 2004). In his work, Kernberg describes three levels of personality organization on a continuum of reality testing. The most intact reality testing is reflected in merely "neurotic" personality organization, while personality disorders such as BPD and schizophrenia represent more serious and most serious problems of integration and reality testing, respectively. The purpose of the following two sections is to highlight how this continuum of reality testing could be expressed in terms of a free energy formulation focused on the relative influence of dominant versus part-object models.

BORDERLINE PERSONALITY DISORDER

A hallmark of the experience of people diagnosed with borderline personality disorder (BPD) is the instability in their perception of self (identity) and others. In psychotherapy, clients diagnosed as BPD may move easily between extremes of idealization and aggression or persecution (these negative responses are often frequent) in their transference responses to psychotherapists, as well as to perceptions of other people in general (Yeomans et al., 2015).

Figure 3 is again adapted from a figure found by Tretter and Löffler-Stastka (2018). It presents a development of affectively organized object representation in both typical developed configurations and BPD configuration. In the first infant stage, the system's current state (represented by the ball) can more easily move between extreme positive and negative basins of attraction, formalizing a state of instability within this "semi-quantitative" model. The deepening of these basins during development reflects the tendency toward greater affective *stability* (not intensity), with a reduced tendency to move toward opposite poles. The growth of the central barrier through development (marked with vertical lines) could be described as formalizing

[12]While the difference between BPD and schizophrenia here is given as involving different levels of functional connectivity, the specific patterns of connectivity implicated in both may be distinct. This distinction may be an important area for future research.

[13]Though not addressed in this paper, states of reduced consciousness (e.g. sleep, intoxication, fatigue and others) must also reduce the functioning of the dominant platonic model, thereby increasing part-object influence on consciousness.

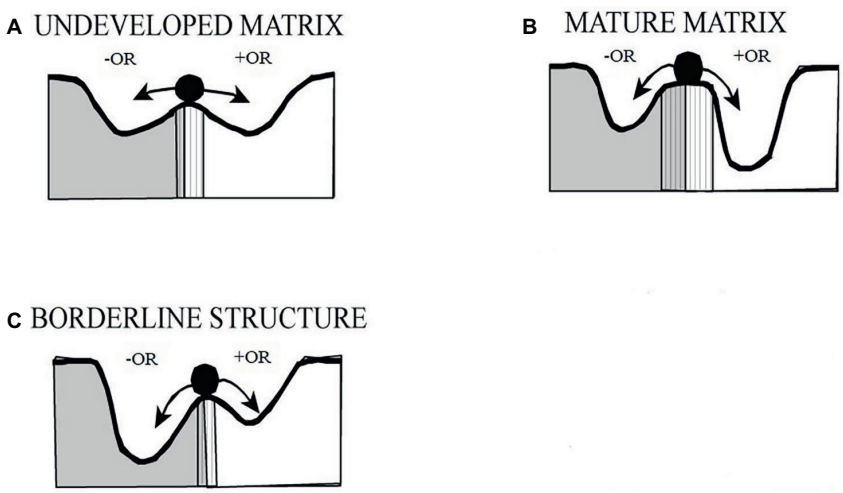

FIGURE 3 | Borderline dynamics of affectively organized object representation (OR), modified from Tretter and Löffler-Stastka, 2018, p. 11, with permission from the original author and copyright holder, Thieme Publishing. These figures show basins of attraction for positive and negative affect for typical early development **(A)**, typical mature development **(B)**, and dynamics of Borderline structure **(C)**.

the increasing influence of dominant generative models of self (Tretter and Löffler-Stastka focus on self-representation with regard to BPD) and of others – the dominant platonic person models described here. As these dominant models begin to grow in influence over the system state, there is a reduction in the tendency to move toward extremes of positive and negative affectively organized states, or between them. Tretter and Löffler-Stastka described BPD as an intermediate position where the boundary between states is less developed (and negative states are a stronger attractor than positive ones).

Their model focuses on "object-related" self-presentation, and while the present paper has not addressed the dynamics of self-representation, the theoretical account that has been put forward in this paper is compatible with the formal account found in their article. Whether it may be due to a predisposition toward subtle problems of functional connectivity that limit the development of highly complex person models[14], or due to an excess of negative affective experience during early development that similarly places constraints on the functional connectivity underwriting the potential complexity of our platonic person model[15] (or both), we may suggest that the dominant platonic person model is less well developed, and less complex, and ultimately less able to entrain (predominantly negative) part-object models in conscious perception of other people.

SCHIZOPHRENIA

Impairment in reality testing of perception or beliefs is the key defining characteristic of psychotic disorders such as schizophrenia. This often manifests in delusional thought content, which appears relatively impervious to any contradictory information, particularly persecutory delusion (American Psychiatric Association, 2013). While people diagnosed with schizophrenia may have a variety of delusional thoughts over their lifespan, there is often a core (typically persecutory) delusion that never shifts, even among those who successfully maintain a residual phase following some number of breakdowns, though it may reduce in importance to the thought process of those who are relatively well. These core persecutory delusions may often be bizarre, such as perceptions that other people are demons or witches, or similar. These symptoms may occur against a backdrop of a relative poverty of thought, particularly in chronic cases with a history of frequent hospitalizations.

Computational approaches to neuropsychiatry such as those using a free energy principle framework have approached schizophrenic symptoms as rooted in a deficit of functional connectivity and hence of complex generative models (Montague et al., 2012) and have approached persecutory delusion (PD) as aberrant social perception related to impairment of generative models of others (Diaconescu et al., 2019). These accounts may offer a satisfying account of the failure of more realistic perception, and Friston et al. (2016) suggested that the relative persistence of false beliefs (delusion) in schizophrenia reflects an increased precision given to prior (false) beliefs in response to failures of attenuation of sensory information. However, what their account does not clarify is what sorts of inferences are like to come to the fore in PDs, or rather what the reason is for the specific affective or thematic nature of those false beliefs. Certainly, the schizophrenic person's inferences that external influences are controlling their experience and behavior as described in their paper are likely to give rise to negative affect and inferences of persecution, which is of course possible. However, the present paper offers an alternative suggestion in which the specific affective and thematic characteristics of the delusional experience are the consequence of *pre-existing* models, which offer the best active inference for the abnormal conditions present in the brain.

An established idea in psychoanalysis is that psychotic experiences of PD may in some sense be founded on split-off persecutory bad objects in the Kleinian sense (Klein, 1946; Segal, 1964). Within the current formulation, we might suggest that the failure to maintain a highly developed platonic person model that regulates free energy in daily encounters, the system falls back on less well developed, but meaningfully established part-object models (such as a persecutory bad object) that require far less effective connectivity to operate.

Though these part models are far less able to reduce the free energy of the system (they integrate far less sensory input than typical dominant models do), in a sense they become "the only game in town" as the only regulatory generative model available that can explain away the person's social experience. This may go some way to explaining the relative intransigence of such core PDs (formally described as increased precisions of these deep priors), as they become the central foundation of the person's social perception[16] and even start to undergo some development and updating themselves (e.g. the patient forms detailed verbal structures around them), though this is clearly limited by the general constraints offered by the problems in connectivity.

This also goes some way to explaining their persecutory character. We might explain this in a narrative way. First,

[14]Witt et al. (2017) have shown how BPD has some genetic overlap with Bipolar Disorder, Major Depressive Disorder, and Schizophrenia, including genes implicated in cell adhesion and myelination: "The gene-set analyses yielded significant results for exocytosis. In neuronal synapses, exocytosis is triggered by an influx of calcium and critically underlies synaptic signaling. Dysregulated neuronal signaling and exocytosis are core features of neurodevelopmental psychiatric disorders such as the autism spectrum disorders and intellectual disability" (p. 5). It seems possible that these genetic differences all contribute to limits on functional connectivity in various ways that in turn place limits on the potential complexity of generative models that may regulate affective functioning.

[15]Duque-Alarcón et al. (2019) reported finding atypical brain functional connectivity in BPD patients who have experienced childhood maltreatment.

[16]The author worked for several years in a community-based residence for people suffering from psychotic disorders and gained the impression that the patients' relationship to their core delusion seemed formally similar to an attachment, due to the great importance given to the delusional content, how it was often invoked when dealing with stressors, the distress experienced when it was threatened by information or insensitive interaction in some way, and the efforts to reject or avoid situations or people who threatened these perceptions. The reason being given in this paper is that that delusional content has become the central foundation of the conscious self in many of these patients.

during the long, distressing prodromal period, the inability of the dominant platonic model (and perhaps self- and various other models as well) to integrate experience[17] leads to increasing free energy. This progressive failure of the dominant model and escalation of free energy and negative affect likely activates negative affective part objects (which form in similar circumstances early in our development) in a feedback loop, and they begin to gain dominance in our conscious perception as they increasingly become the best (or only) available inference about our (social) experience. As the person reaches the more acute phase, the increasing failure of dominant models and ascendance of part-object perception drives the magnitude of the derealization described earlier in the paper (in the "extreme" example of the man who sees his ex-partner with a new lover), an overall situation which is all the more fundamentally traumatic as it does not go away after a short time.

The present formulation of psychotic phenomena supports the psychoanalytic perspective of the compensatory or defensive nature of the positive symptoms of psychosis, first articulated by Freud in "The Neuro-Psychoses of Defence" (Freud, 1894/1962) and now supported by the description by Friston et al. (2016).

The remainder of this paper is given over to a discussion of the implications and contributions of the current paper, as well as the problem of evidence of the current formulation as well as possibilities for future research.

CONTRIBUTION AND IMPLICATIONS

In order to make sense of any implications of the current paper, it is necessary to clarify what the specific unique contribution it aims to make to existing literature. The present paper is offered as theory. In doing so, it builds on existing theory. The theory put forward in this paper is built on ideas within object relations psychoanalysis, specifically on Klein's (1946) theory of splitting, good and bad part objects and whole object integration. It is also built on Kernberg's (1965, 1987) description of realistic and fantasy components of object representations and makes use of his (Kernberg, 1984, 1996, 2004) work on a continuum of levels of reality testing through neurotic, personality disordered and schizophrenic personality organization. What is uniquely offered here is an attempt to state these theories in such a way that they fit within a newer theoretical paradigm, which could broadly be subsumed under the free energy principle. In this way, it is built on Friston's (2010) theory of the free energy principle as well. So, the uniqueness lies in the attempt to marry these theories.

The value of this union from the perspective of psychoanalysis could be said to be twofold. First, the value of stating the psychoanalytic theory in terms of Friston's free energy principle (FEP) theory lies in the value of the FEP theory itself and the contribution it brings to psychoanalytic thinking. Second, the value of stating the psychoanalytic theory in this way also lies in connecting it to a broader system-theory perspective of the world, within which the FEP theory could be said to reside as well.

First, the value that comes from the FEP is that it articulates a neuronally plausible process theory regarding regulation and message passing within the nervous system (Friston et al., 2017) that offers anatomical constraints on those processes, which can and have been assessed empirically (Parr and Friston, 2018). This allows a stronger explanation of the phenomena described by psychoanalytic theories than the psychoanalytic theories themselves, which have historically not been adequately connected to neurophysiological processes, nor even to other psychoanalytic ideas. This lack of a functional psychoanalytic metapsychology can be demonstrated by asking the question: why do objects form in the psyche? If we follow Freud's partially failed energic explanation (Connolly, 2016, unpublished) we could say objects form because they bind free-floating energy, but we would be unable to adequately link Freud's energy with neurophysiological process. If we take Klein's (1946) suggestion that the formation of objects manages anxiety, then we link it to an abstract emotional construct, but not to neurophysiological process. But if we follow a free energy explanation, we say that objects form because they minimize free energy, through maximizing accuracy of generative models as parsimoniously as possible. We are then on firmer ground, as this explanation is rooted on the neuronal foundation described by Friston's (2010) work.

The field that is expanding around the free energy principle is itself embedded within a broader framework that is well described as systems theory, which is the second benefit of the union of theories offered in this paper. A fuller description of systems theory and its potential value to psychoanalysis can be found by Connolly and van Deventer (2017) but can be heavily summarized here as saying that a system view of the world is a hierarchical one, where system is superimposed on system and so on.

This hierarchical perspective can be seen in the view expressed by Tretter and Löffler-Stastka (2018) when they call for an integrative clinical systems psychology:

> "… The crucial term 'system' is defined as a set of elements and a set of relations (structure and connectivity), … In line with this definition, a system can be characterized simply by the term 'structure' or by the popular expression 'network' (nodes and edges) as it is a network with boundaries. Or, with other words: a living system is a network (or structure) with boundaries. Properties of systems are states (e.g., equilibrium, non-equilibrium) and processes, some of them have goal-directed functions as a subset of activities. … Systemic exploratory methodology basically implies to zoom into the micro-level of the

[17]This may be due to some form of progressive deterioration of the dominant model(s) or perhaps to an increase in demands on the person that exposes an existing fragility of the dominant models. Discussions with male patients often revealed that their first acute episodes happened together with an increased demand for autonomy and reduction of parental care, such as entering the army or going away to study. However, this idea seems more difficult to apply to the typically older onset for female patients.

> subject of study, not forgetting the context and also to zoom out to the macro-level without forgetting the details. If we zoom out of the detailed consideration of elementary functions of the mind to a more holistic view we will refer to several holistic models that also will provide a diversified understanding of mental processes in context of clinical issues" (p. 7).

Their work suggests that we might define a system as a set of elements and relations between them, where if we "zoom in" to higher resolution we see that each of those elements is itself composed of a system of elements and relations, and so on. The key point they make is that most major theories of psychology might be represented in an abstract description of this form, where theories are not "floating" in an abstract space where they merely have a heuristic or *ad hoc* role in explaining research findings but rather are embedded in a larger superstructure. In this way, different theories (including at different levels of organization in the person) can be integrated with one another. This offers the hope of convergence in our theoretical work, rather than the seemingly endless divergence of theory that has taken place in the field of psychology.

The free energy principle and the body of theory that are growing under its ambit fit the bill of a system-based theory that offers clear system principles and a basis for hierarchical organization of systems and sub-systems. The FEP paradigm can "zoom down" to show how the FEP-based organization of living systems is founded upon inorganic processes (Friston, 2019, submitted) and equally, zoom up to social, cultural, and environmental systems that entrain living systems (Connolly and van Deventer, 2017; Badcock et al., 2019).

Specifically, in this paper, this hierarchical embeddedness of the processes described lies in the foundation of the affective systems described by Panksepp (1998) and how their cortical influence in the form of part objects steadily becomes entrained by a history of social interaction, which comes to form the dominant platonic object.

This integration of the theory of objects and part objects with a system-based FEP perspective now also allows integration with the psychoanalytic principle of conflict, which was integrated with a FEP perspective in the work by Hopkins (2016) and Connolly (2018). This has allowed the current paper to offer a conceptual account of how conflict can lead to the splitting off of part objects and thereby integrating these different psychoanalytic theories rather than leaving them separated across the gulf of their respective Freudian and Kleinian paradigms. Through a steady work of application of system-based ideas in this way, a new psychoanalytic model of the mind may eventually emerge.

Beyond these very broad implications, the integration with a free energy principle account has more specific implications for how we conceive of objects. Some of these are highlighted next:

1. A part object is here described as a generative model. This means that it reflects a distinct anatomical expression with a Markov blanket. This itself has a number of implications. One key one here is that it "tries" to maintain its own existence and avoid destruction (phase change). In other words, one could state it intuitively as saying that the object has a "life of its own." This also means seeking to accumulate evidence for its own existence. This supports Freud's (1912/1963) idea that we appear to seek transferences out (try to apply them to each new person we meet).
2. Part objects must have some success in predicting situations, or people's behavior, or they could never be sustained. This might explain the common preference for entertainment that portrays people in "archetypal" ways. In this way, part objects can accumulate evidence. This would also be true for a common preference to "want" to see others in distorted ways, for example, seeming to "relish" describing someone as a villainous person.
3. While part objects may be "starved" somewhat, in the sense of being prevented from accumulating evidence in some way, they are difficult to get rid of, for the reasons indicated in the previous points. However, they may be entrained, which essentially means being increasingly merged with a more dominant, integrated model. Practically, this could mean the further development of the dominant model (such as through mentalization), as well as recognition, insight, and perhaps also acceptance of these relevant qualities in oneself and others.
4. Recently, Ramstead et al. (2019, submitted) argued that hierarchical generative models do not so much have the characteristics of representation as they do of control. That means that part objects, as well as dominant objects, are not just representations but rather realize the function of control in the psyche and integrate relevant actions in a sense as well. As Ramstead et al. (2019, submitted) suggested: "… 'perceptual inference' is just one moment of the policy selection process in active inference under the FEP, namely, state estimation. The issue we want to press here is that the active inference framework implies that perception is a form of action, that is, action and perception cannot be pulled apart …" (p. 2). This means that part objects are perhaps best not thought of just as representations of perceptual memories, unless we think of memories as control mechanisms in the same way as well.

These potential implications are just a beginning, and further implications may be uncovered with further progress.

While the integration of these psychoanalytic theories with the free energy paradigm has many tangible benefits for the body of psychoanalytic theory, the question might well be asked what they offer to the growing field within the active inference and the free energy. As stated earlier in this paper, the central value of the psychoanalytic literature is a long history of observations and clinical insights that can help direct research. In this case, it may generate interest in research into the role of part objects of the kind described here, in perception.

EVIDENCE AND FUTURE RESEARCH

A critical problem with the current paper is the lack of empirical evidence for its central claims. The central claims are as follows:

1. Part-object models organized by affect typically exist in human nervous systems since early childhood.
2. They may sometimes not be entrained by a dominant model (perhaps due to conflict), and a competitive relationship may exist between such split off part-object models and dominant ones in order to determine the process of active inference.
3. Increasing levels of influence of part objects (and corresponding decreases of influence of dominant models) on a continuum from transient emotions, to personality disorder (e.g., BPD) and to schizophrenia, in order, probably due to problems in connectivity which underlie dominant models.

As such, none of the research referred to in this paper directly proves these core hypotheses.

Rather, the present paper has taken the form of an argument and has used research findings along the way to support specific points and assumptions being made during its course. For example, the claim that affect may play a foundational role to object formation is supported by making reference to Panksepp's (1998) work on affective command systems. Claims regarding the role of connectivity in reality testing were supported with empirical findings regarding connectivity in transient emotional experiences (Eryilmaz et al., 2011), borderline personality disorder in terms of genetic predisposition (Witt et al., 2017) as well as early experiences of distress (Duque-Alarcón et al., 2019), and in schizophrenia (Friston et al., 2016). Evidence for hierarchical layers of processing in social inference and theory of mind was offered from the work of Diaconescu et al. (2017, submitted).

This kind of "amalgamation" of different sources of contributory evidence does not constitute proof of a theory but may be a critical for development as well as refinement of theory (Fletcher et al., 2019; Kao, 2019). This form of evidence can suggest that a theory is plausible rather than confirm it. In turn, plausibility is an important guide to which theories should be investigated further, and which not (Bertolaso and Sterpetti, 2019).

This form of amalgamation of evidence may be unavoidable when faced with theories that are difficult to prove:

> "When access to phenomena of interest is incomplete, piecemeal, indirect, or mediated by substantial auxiliary assumptions, it is not always obvious in what manner scientists can justifiably decide how their total evidence comparatively supports hypotheses and informs future research" (Fletcher et al., 2019, p. 3164).

In this case, the challenge is presented by the likelihood that both part objects and objects are encoded in complex multiple areas of the cortex and involve multi-level processes that unfold over time. This makes it more than challenging to isolate specific objects in brain-imaging research. This challenge can be seen more clearly when one tries to locate the part- and dominant-object models in Panksepp's scheme of emotions, the primary, secondary, and tertiary emotions. At their outset, when part objects (and the beginnings of the dominant model) form, they fit most closely with the secondary layer described by Panksepp (2010), in which they are shaped by basic learning processes not dependent on any tertiary-level processes in the beginning. However, if we try find some consilience between the tertiary-level processes described by Panksepp on the one hand, and the consideration of alternative policies of action that have reached sufficient "temporal thickness" or "counterfactual depth" (which Friston, 2018, described as foundational to consciousness) on the other, we could say that both part-object models and dominant models are reflected in tertiary level processes as well (though the dominant ones usually much more so). Clearly, both must involve some encoding at a cortical level, though with dominant models probably reflected by more connections and distribution than part-object ones.

In this way, cortical representation of long-term memory must play a role in the formation of platonic models. While it has been suggested above that part-object models are more than just memory representations of perceptual experiences, action selection is an inherent aspect of working memory, which activates those representations. In his paper "Cortex and Memory: Emergence of a New Paradigm," Fuster (2009) describes a situation demonstrating this difficulty with regard to long-term memory networks, which become activated in working memory:

> "… [A] memory or an item of knowledge consists of a widespread cortical network of connections, formed by experience, that joins dispersed cell populations. … A complex memory network, … is largely interregional, linking neuron assemblies and smaller networks in separate and noncontiguous areas of the cortex" (p. 2048).

These challenges do not mean that proof is impossible. However, the requirement in this case would require brain imaging data that compare transient states such as in intense emotions conceptually related to part objects, with longitudinally obtained data of brain states in early childhood, to say if they are similar. This is of course made difficult due to the changes that occur in maturation.

In the absence of such evidence, system models of this kind often make use of different strategy that involves simulation and application of mathematical modeling.

> "… [W]e start with verbal models that explicate interactions and that in some cases are presented in graphs. Usually the next step should be a mathematical formalization of this hypothetical causal model but we don't think this will really increase evidence here and therefore it should be reserved for a later step of discussions of modeling the mind. After the formalization, empirical data should be integrated and now it is possible to transform the model to a computer algebra system (e.g., Maple R, Matlab R, Mathematica R) for running

simulations in order to explore the functional structure of the model by process analysis. This stepwise procedure was developed basically in the context of systems dynamics ….” (Tretter and Löffler-Stastka, 2018).

The study by Moutoussis et al. (2014b) is an example of such application of a mathematical model applied to a simulation, and the results compared with what is expected. It is hoped that the present work might stimulate further research of a similar kind, which may model the relative influence of part object and dominant models of people.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

The author thanks the reviewers for their helpful remarks and additions and the author is especially grateful to Dr Jeremy Holmes, who read and commented on an earlier version of this paper.

REFERENCES

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders*. 5th Edn. Washington, DC: American Psychiatric Association.

Badcock, P. B., Friston, K. J., and Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: a free-energy formulation of the human psyche. *Phys. Life Rev.* pii: S1571-0645(19)30002-8. doi: 10.1016/j.plrev.2018.10.002 [Epub ahead of print].

Banavar, J. R., Maritan, A., and Volkov, I. (2010). Applications of the principle of maximum entropy: from physics to ecology. *J. Phys. Condens. Matter* 22:063101. doi: 10.1088/0953-8984/22/6/063101

Barrett, L. F., Lindquist, K. A., Bliss-Moreau, E., Duncan, S., Gendron, M., Mize, J., et al. (2007). Of mice and men: natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard. *Perspect. Psychol. Sci.* 2, 297–312. doi: 10.1111/j.1745-6916.2007.00046.x

Bertolaso, M., and Sterpetti, F. (2019). Evidence amalgamation, plausibility, and cancer research. *Synthese* 196, 3279–3317. doi: 10.1007/s11229-017-1591-9

Connolly, P. (2018). Expected free energy formalizes conflict underlying defense in Freudian psychoanalysis. *Front. Psychol.* 9:1624. doi: 10.3389/fpsyg.2018.01264

Connolly, P. (2019). “Reality is hierarchically organized: the recursive foundations of living systems and beyond” in *Focus on systems theory research*. eds. M. F. Casanova and I. Opris (New York: Nova Science Publishers).

Connolly, P., and van Deventer, V. (2017). Hierarchical recursive organization and the free energy principle: from biological self-organization to the psychoanalytic mind. *Front. Psychol.* 8:1695. doi: 10.3389/fpsyg.2017.01695

Diaconescu, A. O., Wellstein, K. V., Kasper, L., Mathys, C., and Stephan, K. E. (2019). *Hierarchical Bayesian models of social inference for probing persecutory delusional ideation*. Poster session presented at the Annual Meeting of the Organization for Human Brain Mapping, Rome, Italy.

Duque-Alarcón, X., Alcalá-Lozano, R., González-Olvera, J. J., Garza-Villarreal, E. A., and Pellicer, F. (2019). Effects of childhood maltreatment on social cognition and brain functional connectivity in borderline personality disorder patients. *Front. Psych.* 10:156. doi: 10.3389/fpsyt.2019.00156

Eryilmaz, H., Van De Ville, D., Schwartz, S., and Vuilleumier, P. (2011). Impact of transient emotions on functional connectivity during subsequent resting state: a wavelet correlation approach. *NeuroImage* 54, 2481–2491. doi: 10.1016/j.neuroimage.2010.10.021

Fletcher, S., Landes, J., and Poellinger, R. (2019). Evidence amalgamation in the sciences: an introduction. *Synthese* 196, 3163–3188. doi: 10.1007/s11229-018-1840-6

Freud, S. (1894/1962). “The neuro-psychoses of defence” in *The standard edition of the complete psychological works of Sigmund Freud, volume III (1893–1899): Early psycho-analytic publications*. ed. J. Strachey (London: Hogarth Press), 41–61.

Freud, S. (1912/1963). “The dynamics of the transference” in *The collected papers of Sigmund Freud: Therapy and technique*. ed. P. Rieff (New York: Collier Books), 105–115.

Freud, S. (1915/1963). “Repression” in *The collected papers of Sigmund Freud volume 6: General psychological theory: Papers on metapsychology*. ed. P. Rieff (New York: Collier Books), 104–115.

Freud, S. (1919/1955). “The ‘uncanny’” in *The standard edition of the complete psychological works of Sigmund Freud, volume XVII (1917–1919): An infantile neurosis and other works*. ed. J. Strachey (London: Hogarth Press), 217–256.

Friston, K. J. (2010). A free energy principle for the brain. *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J. (2018). Am I self-conscious? (or does self-organization entail self-consciousness?). *Front. Psychol.* 9:579. doi: 10.3389/fpsyg.2018.00579

Friston, K. J., Brown, H. R., Siemerkus, J., and Stephan, K. E. (2016). The dysconnection hypothesis. *Schizophr. Res.* 176, 83–94. doi: 10.1016/j.schres.2016.07.014

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Fuster, J. M. (2009). Cortex and memory: emergence of a new paradigm. *J. Cogn. Neurosci.* 21, 2047–2072. doi: 10.1162/jocn.2009.21280

Gershman, S. J., and Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* 20, 251–256. doi: 10.1016/j.conb.2010.02.008

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature* 436, 801–806. doi: 10.1038/nature03721

Haken, H. (1983/2004). *Synergetics: Introduction and advanced topics*. Berlin: Springer-Verlag. Original work published in 1983.

Hobson, J. A., Hong, C. C., and Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Front. Psychol.* 5:1133. doi: 10.3389/fpsyg.2014.01133

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Hopkins, J. (2012). “Psychoanalysis, representation and neuroscience: the Freudian unconscious and the Bayesian brain” in *From the couch to the lab: Psychoanalysis, neuroscience and cognitive psychology in dialogue*. eds. A. Fotopoulu, D. Pfaff, and M. Conway (Oxford: Oxford University Press), 230–265.

Hopkins, J. (2016). Free energy and virtual reality in neuroscience and neuropsychoanalysis: a complexity theory of dreaming and mental disorder. *Front. Psychol.* 7:922. doi: 10.3389/fpsyg.2016.00922

Isomura, T., and Friston, K. (2018). In vitro neural networks minimise variational free energy. *Sci. Rep.* 8:16926. doi: 10.1038/s41598-018-35221-w

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.

Kao, M. (2019). Unification beyond justification: a strategy for theory development. *Synthese* 196, 3263–3278. doi: 10.1007/s11229-017-1515-8

Kernberg, O. (1965). Notes on countertransference. *J. Am. Psychoanal. Assoc.* 13, 38–56.

Kernberg, O. F. (1976). *Object relations theory and clinical psychoanalysis*. Lanham, MD: Jason Aronson.

Kernberg, O. F. (1984). *Severe personality disorders: Psychotherapeutic strategies*. New Haven: Yale University Press.

Kernberg, O. F. (1987). Projection and projective identification: developmental and clinical aspects. *J. Am. Psychoanal. Assoc.* 35, 795–819.

Kernberg, O. F. (1996). “A psychoanalytic theory of personality disorders” in *Major theories of personality disorder*. eds. J. F. Clarkin and M. F. Lenzenweger (New York: Guilford Press).

Kernberg, O. F. (2004). “Borderline personality disorder and borderline personality organization: psychopathology and psychotherapy” in *Handbook of personality disorders. Theory and practice*. ed. J. J. Magnavita (New York: Wiley), 92–119.

Klein, M. (1946). Notes on some schizoid mechanisms. *Int. J. Psychoanal.* 27, 99–110.

Lamm, C., Bukowski, H., and Silani, G. (2016). From shared to distinct self-other representations in empathy: evidence from neurotypical function and

socio-cognitive disorders. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 371:20150083. doi: 10.1098/rstb.2015.0083

Maisto, D., Donnarumma, F., and Pezzulo, G. (2015). Divide et impera: subgoaling reduces the complexity of probabilistic inference and problem solving. *J. R. Soc. Interface* 12:20141335. doi: 10.1098/rsif.2014.1335

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80. doi: 10.1016/j.tics.2011.11.018

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., and Friston, K. J. (2014a). Bayesian inferences about the self (and others): a review. *Conscious. Cogn.* 25C, 67–76. doi: 10.1016/j.concog.2014.01.009

Moutoussis, M., Trujillo-Barreto, N. J., El-Deredy, W., Dolan, R. J., and Friston, K. J. (2014b). A formal model of interpersonal inference. *Front. Hum. Neurosci.* 8:160. doi: 10.3389/fnhum.2014.00160

Ogden, T. H. (1992). The dialectically constituted/decentred subject of psychoanalysis. II. The contributions of Klein and Winnicott. *Int. J. Psychoanal.* 73, 613–626.

Oosterwijk, S., Lindquist, K. A., Anderson, E., Dautoff, R., Moriguchi, Y., and Barrett, L. F. (2012). States of mind: emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage* 62, 2110–2128. doi: 10.1016/j.neuroimage.2012.05.079

Panksepp, J. (1998). *Affective neuroscience*. Oxford: Oxford University Press.

Panksepp, J. (2010). Affective neuroscience of the emotional BrainMind: evolutionary perspectives and implications for understanding depression. *Dialogues Clin. Neurosci.* 12, 533–545.

Parr, T., and Friston, K. J. (2018). The anatomy of inference: generative models and brain structure. *Front. Comput. Neurosci.* 12:90. doi: 10.3389/fncom.2018.00090

Paton, B., Hohwy, J., and Enticott, P. G. (2012). The rubber hand illusion reveals proprioceptive and sensorimotor differences in autism spectrum disorders. *J. Autism Dev. Disord.* 42, 1870–1883. doi: 10.1007/s10803-011-1430-7

Segal, H. (1964). *Introduction to the work of Melanie Klein*. New York: Basic Books.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* 17, 565–573. doi: 10.1016/j.tics.2013.09.007

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., and Frith, C. D. (2004). Empathy for pain involves the affective but not the sensory components of pain. *Science* 303, 1157–1161. doi: 10.1126/science.1093535

Tervo, D. G., Tenenbaum, J. B., and Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Curr. Opin. Neurobiol.* 37, 99–105. doi: 10.1016/j.conb.2016.01.014

Tretter, F., and Löffler-Stastka, H. (2018). Steps toward an integrative clinical systems psychology. *Front. Psychol.* 9:1616. doi: 10.3389/fpsyg.2018.01616

Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M. B., et al. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature* 561, 57–62. doi: 10.1038/s41586-018-0459-6

Witt, S. H., Streit, F., Jungkunz, M., Frank, J., Awasthi, S., Reinbold, C. S., et al. (2017). Genome-wide association study of borderline personality disorder reveals genetic overlap with bipolar disorder, major depression and schizophrenia. *Transl. Psychiatry* 7:e1155. doi: 10.1038/tp.2017.115

Yeomans, F. E., Clarkin, J. F., and Kernberg, O. F. (2015). *Transference-focused psychotherapy for borderline personality disorder: A clinical guide*. Arlington, VA, US: American Psychiatric Publishing, Inc.

Zeller, D., Litvak, V., Friston, K. J., and Classen, J. (2015). Sensory processing and the rubber hand illusion--an evoked potentials study. *J. Cogn. Neurosci.* 27, 573–582. doi: 10.1162/jocn_a_00705

Conflict of Interest: 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 Psychology

REVIEW
published: 28 February 2020
doi: 10.3389/fpsyg.2020.00325

From the Principle of Inertia to the Death Drive: The Influence of the Second Law of Thermodynamics on the Freudian Theory of the Psychical Apparatus

Jessica Tran The[1,2,3,4]*, *Jean-Philippe Ansermet*[5], *Pierre Magistretti*[3,6,7] and *François Ansermet*[1,3,8]

[1] Faculty of Biology and Medicine, Université de Lausanne, Lausanne, Switzerland, [2] Département D'études Psychanalytiques - UFR IHSS, Université de Paris, Paris, France, [3] Agalma Foundation Geneva, Geneva, Switzerland, [4] Centre de Recherches Psychanalyse, Médecine et Société, Université Paris Diderot, Paris, France, [5] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [6] Institut de L'esprit de Cerveau, École Polytechnique Fédérale de Lausanne, Sion, Switzerland, [7] Lausanne University Hospital (CHUV), Lausanne, Switzerland, [8] Faculté de Médecine, Département de Psychiatrie, Université de Genève, Geneva, Switzerland

OPEN ACCESS

Edited by:
Christoph Mathys,
International School for Advanced Studies (SISSA), Italy

Reviewed by:
Karl Friston,
University College London, United Kingdom
Tobias Nolte,
University College London, United Kingdom

***Correspondence:**
Jessica Tran The
jessica.tranthe@ens.fr

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: 30 October 2019
Accepted: 11 February 2020
Published: 28 February 2020

Citation:
Tran The J, Ansermet J-P, Magistretti P and Ansermet F (2020) From the Principle of Inertia to the Death Drive: The Influence of the Second Law of Thermodynamics on the Freudian Theory of the Psychical Apparatus. Front. Psychol. 11:325. doi: 10.3389/fpsyg.2020.00325

In the Freudian theory of the psychical apparatus, the introduction from the 1920s onward of the second drive dualism appears as a major turning point. The idea of a "death drive," first expressed in *Beyond the Pleasure Principle* (Freud, 1920), is generally considered to be a new concept, one that represents a break with Freud's previous thinking. It has often surprised the scholars because it seemed, at first sight, difficult to reconcile with the idea of the singularity of living organisms within which the psychical functions form an integral part. Our research aims to demonstrate that the theory of the death drive does not represent a complete change in direction for Freud. It is present, in essence, in his earliest work, to the extent that the "principle of inertia" described in 1895 in *A Project for a Scientific Psychology* (Freud, 1895) can be seen as a precursor to the death drive. Based on a reading of Freud's early formulations of his ideas, we aim to bring to light how certain aporias that seem inherent to the concept of the death drive can be overcome if we consider them in the context of an epistemological model that draws on the paradigms of physics which were conveyed by the Helmholtz School. Namely, we can consider the idea of death drive in reference to the principle of entropy and the laws of thermodynamics.

Keywords: Freud, principle of inertia, death drive, thermodynamics, entropy

INTRODUCTION

The establishment of the idea of the "death drive" at the theoretical turning point of the 1920s is often seen as a significant watershed in Freudian theory. A defining moment, that led to profound epistemological reconfigurations, with the advent of a new dualism of the drive. Indeed, in his essay *Beyond the Pleasure Principle* (Freud, 1920) Freud introduces a distinctly subversive theory. If, in

his definition of the pleasure principle, he had recognized that "the effort to reduce, to keep constant or to remove internal tension" (Freud, 1920, p. 55) constitutes "The dominating tendency of mental life, and perhaps of the nervous life in general" (Freud, 1920, p. 55), he made this concept even more radical by introducing an apparently paradoxical theory:

> *"the life process of the individual leads for internal reasons to an abolition of the chemical tension, that is to say, to death." (Freud, 1920, p. 55)*

If this concept can, in some respects, appear to be resolutely surprising, and especially novel in Freudian thinking, a return to Freud's first theoretical ideas – in particular *A Project for a Scientific Psychology* (Freud, 1895) – shows that this concept existed already in essence as far back as his first theoretical elaborations; most notably in the description of the "principle of inertia." Indeed, Freud had, as early as 1895, postulated the existence of a first principle within the functioning of the nervous system, a principle that would consist in trying to achieve a "zero level" of excitation: the principle of neuronal inertia is defined as the fact that "neurons tend to divest themselves of (Q)" (Freud, 1895, p. 296). The "quantity (Q)" here stands for the quantity of neuronal excitation, a theory that constitutes the "first fundamental definition" of *A Project for a Scientific Psychology* (Freud, 1895). Thus, it is not a principle of constancy, or of homeostasis, that would hark back to the general tendency of the organism to maintain a positive optimum level (be it in body temperature, concentration of oxygen in the blood, etc.), that Freud places at the basis of the psychical function. On the contrary, he defines the original tendency of the neuronal system as one of trying to reach a "level = 0"; the equivalent of a search for a total absence of excitation, or of the fastest possible discharge of the Q quantities to re-establish a level of zero (Freud, 1895). This hypothesis at the root of the whole Freudian theoretical construction can, in a way, if we follow the arguments of Laplanche and Pontalis "appear to be an aberration from the point of view of the life sciences" (Laplanche and Pontalis, 1967a, p. 326), this in so far as that it postulates a first principle in the functioning of the nervous system which would be "the negation of any stable difference in level" (Laplanche and Pontalis, 1967a, p. 326). Would psychoanalysis therefore be, in its fundamentals, radically estranged from a theory of the living organism? To shed light on this aporia it is necessary to put the definition of "principle of inertia" back into its context, within the radically physicalist epistemological model that Freud draws on; Freud being a faithful heir to the tradition of the Hemholtz school, and its contemporaneous research in thermodynamics (Tran The et al., 2018). More specifically, it is important to consider how certain complexities inherent to the definition of the Freudian principle of inertia, and to the death drive as its logical continuation in the 1920s, can be explained through reference to entropy, that is to say the second law of thermodynamics. Thus, we can observe how a return to the physicalist epistemological foundations of the Freudian theoretic model makes it possible, in a surprising way, to lift the apparent aporias between the presence of a death drive at work in the psyche on the one hand, and on the other the indissociable natures of the link between the psychical function and the living organism within which it occurs.

THE INFLUENCE OF THE HELMHOLTZ SCHOOL OF PHYSICS ON FREUD'S TRAINING

In order to understand why the death drive might, on first impression, "seem like an aberration from the point of view of the life sciences" (Laplanche and Pontalis, 1967a, p. 326) (translated for this publication), most notably from the point of view of biology and physiology, we need to remember that the dominant epistemological paradigm during the formative years of Freud's medical training was very different to the "physiological revolution" instigated by Claude Bernard in 1860s France. Bernard's influence had established physiology as an independent and autonomous discipline in relation to physical chemistry (Tran The et al., 2018). However, it is in a radically different geographic and scientific context that Freud undertook his medical studies, at the Vienna Faculty, in autumn 1873 (Jones, 1953). Beyond the Rhine, physiology's autonomy with regards to physics was far from complete, and it was within a paradigm that is profoundly antagonistic with regards to that of French biology, that Freud's training took place. At the end of his third year, Freud joined Ernest Brücke's physiology laboratory. Freud viewed Brücke as a "model" (Freud, 1925). Besides the respect and admiration that Freud felt for this undisputable master, this filiation bears witness to Freud's adherence to a whole scientific paradigm to which he would make himself heir. As Jones (1953) underlines, the Brücke Institute had close ties with the Helmholtz School. The story of this scientific movement began in the 1840s, with the friendship between various physiologists trained in Johannes Müller's theories of specific nerve energies (Assoun, 1981). Du Bois-Reymond, Brücke, Helmholtz, and Ludwig were medical men who appear to have been driven by a veritable "spirit of crusade," who, as Du Bois-Reymond tells it, had undertaken to "pledged a solemn oath to put into effect this truth: "No other forces than the common physical-chemical ones are active within the organism." (Jones, 1953, p. 40). Although they were all medical men by training, their scientific ideas were completely subordinated to the science of physics. This little group, augmented by the arrival of new members, young physics and physiology students who opposed vitalism, became in 1845 the Berliner Physikalische Gesellschaft, the Berlin Physics Society (Jones, 1953). In less than 30 years, they came to dominate the German scientific landscape, becoming the most influential professors of medicine and physiology of the time, and in their turn training a whole generation of students; amongst them Freud and Wundt. Thus, it was a practice characterized by its diversity and its lack of specialization that Freud inherited during his years of training at the Brücke Institute. However, it was physics that represented for all those related disciplines the epistemological model *par excellence*. We can see here that the German school of physiology positioned itself within a movement exactly the reverse of Bernardian physiology. Whereas in France there is a demand for a certain

independence for physiology, as a science in its own right, autonomous from physics. The Berlin medical practitioners sought, on the contrary, to subordinate physiology to physics, making the former an offshoot of the latter. It is then, a physiology radically subordinated to physics as foremost dominant science – to which all natural phenomena must be reduced, including those relating to living organisms – that Freud would make himself heir. If it is within this framework that Freud received his training at the Brücke Institute, the major influence exerted by Helmholtz needs to be underlined. Of all the scientist at the Berliner Physikalische Gesellschaft, Helmholtz was, without doubt, the most eminent. Freud saw him as one of his idols, and would regret all his life not having met him in person (Jones, 1953).

In particular, Helmholtz upheld an understanding of nature in terms of mechanics, and the majority of the physiologists of the powerful German school (Liebig, Ludwig, Müller, Du Bois-Reymond, Virchow, and Brücke) adopted his view according to which "the physical-chemical functioning of the living being is subject to the same laws as inanimate matter, and must be studied on the same terms" (Prigogine and Stengers, 1979) (translated for this publication). It is therefore, under the influence of this theoretical framework of an essentially physical epistemological model (and not physiological in the Bernardian sense) that the first ideas expressed by Freud regarding the principle of inertia would develop, when he was writing *A Project for a Scientific Psychology*. Freud (1895). The radically physicalist scientific environment of his years in medical training would leave an enduring mark on the whole of Freud's corpus. We can see this right up to his later work on the death drive at the turn of the 1920s.

THE PRINCIPLE OF NEURONAL INERTIA IN *A PROJECT FOR A SCIENTIFIC PSYCHOLOGY*

The manuscript of *A Project for a Scientific Psychology* (Freud, 1895), written by Freud in September 1895, can be seen as establishing a continuation with the "Theoretical Considerations" that Breuer had contributed to *Studies on Hysteria* (Freud and Breuer, 1895). This work, that remained unpublished during Freud's lifetime, explicitly expresses the intention to "furnish a psychology that shall be a natural science" (Freud, 1895, p. 295) by describing the psychical processes in terms of "quantitatively determinate states of specifiable material particles" (Freud, 1895, p. 295), neurons – in order to make these processes "perspicuous and free from contradiction" (Freud, 1895, p. 295). Thus, Freud (who like his colleague Breuer views the psychopathological phenomenon of hysteria as an excessive excitation that is impossible to discharge via the usual outlets) will also attempt an explanation in terms of neurophysiology. He does this through a description of the structure and functioning of the nervous system, or "neuronal" system. It was therefore, initially, for the use of neurologists that this project for a scientific psychology was intended. Consequently, Freud retains in his text the energetic and quantitative reference to nervous excitation, which he calls "neuronal excitation"; but abandons the distinction established by Breuer between a "quiescent" energy (the intracerebral tonic excitation), and a "kinetic" energy. This manuscript text, that is clearly neuropsychological in outlook, reveals the first Freudian principles of the regulation of the nervous system. However, when Feud abandons the biological, anatomical, and structural point of view of *A Project for a Scientific Psychology* (Freud, 1895) in favor of the topographical view point of his metapsychology [when he begins Chapter 7 of *The Interpretation of Dreams* (Freud, 1900)], he retains to a great extent his reference to the principles of regulation of the psychical function that he had defined – although their designations will evolve throughout his work.

Committed to the epistemological model of the Helmholtz School, Freud attempts in his project for a scientific psychology to apply the principles of physics to what he terms the "quantity (Q)" of neuronal excitation [fundamental idea of *A Project for a Scientific Psychology* (Freud, 1895)]. This quantity (Q) that according to him is a "quantity in a state of flow": "regarded as (Q), subject to the general laws of motion" (Freud, 1895, p. 295). It is therefore, before any reference to contemporary thermodynamics, primarily to Newton's general theory of motion, that is to classical dynamics, that Freud is referring when he introduces his approach. There exists an obvious intertextuality between the first two parts of the general layout of *A Project for a Scientific Psychology* (Freud, 1895), and the beginning of Newton's *Principia mathematica* (Ansermet, 2019).

In his *The Mathematical Principles of Natural Philosophy* (Newton, 1687/1846), Newton lays down the basis of mechanics, by defining the three laws of motion – a founding act at a turning point in the development of modern science (Ansermet, 2009). Prior to stating his three laws, Newton introduces two fundamental definitions: the first, the "quantity of matter" (Newton, 1687/1846, p. 73), and the second concerning the "quantity of motion" (Newton, 1687/1846, p. 73), are described on the first page of his treatise. In *A Project for a Scientific Psychology* (Freud, 1895), Freud posits as the "first fundamental idea" the concept of "quantity," which corresponds to the neuronal excitation, and defines it from the outset as a "quantity in a state of flow" (Freud, 1895, p. 295). He is, therefore, choosing to "furnish a psychology that shall be a natural science" (Freud, 1895, p. 295) by representing the psychical processes as a "quantitatively determinate states of specifiable material particles" (Freud, 1895, p. 295): neurons. These neurons, which are isolated one from the other, are traversed by quantities of excitation that submit "to the general laws of motion" (Freud, 1895, p. 295). We will recall that as far back as his communication on histology given in 1882, Freud had already defined the neurons as being "isolated routes of conduction" (Freud, 2017).

Having once defined this first fundamental idea of "the quantity of neuronal excitation in motion," Freud goes on to describe a primary and absolutely fundamental principle of the neuronal apparatus. A principle that would regulate the movement of the quantities of excitation (that is to say, their circulation, or their flow, in the neuronal apparatus): the "principle of inertia of neurons." The term "inertia" is totally new to Freud's writings, in so far as that it does not appear

either in *Studies on Hysteria* (Freud and Breuer, 1895), or in his correspondence with Fliess. Henceforth, the principle of neuronal inertia is defined in these terms, "neurons tend to divest themselves of *Q*" (Freud, 1895, p. 296).

When Freud introduces the second fundamental idea of *A Project for a Scientific Psychology* (Freud, 1895), the "theory of neurons," he will seek to combine what he terms a "theory of quantity" – such that the quantity of neuronal excitation is in motion, and thus regulated by the "general laws of motion" – with his knowledge of neurons as he had observed them in the course of his research in histology at the Brücke Institute. Thus, to these considerations coming from physics are added some anatomical views, seen by some as hypothetical, on the structure and the functioning of the nervous system: "we arrive at the idea of a 'cathected' neuron (*N*) filled with a certain quantity (*Qi*), though at other times it may be empty." (Freud, 1895, p. 298). Neurons could therefore be traversed by some form of "current," and be, in accordance with the view already expressed in 1882, "routes for conduction." From there Freud postulates the existence of two types of neuron, sensory neurons and motor neurons, that would enable the nervous system to counteract the reception of quantities by getting rid of them through a reflex motion that discharges the quantity of excitation. It is important to underlined that, contrary to Breuer, Freud does not therefore propose a principle of constancy as the primary tendency of the nervous system. Similarly, he does not refer back to a general tendency of the organism to maintain a positive optimum level (be it body temperature, the concentration of oxygen in the blood, etc.), rather he defines the primal tendency of the neuronal system as the search for a "level = 0" (Freud, 1895). Whereas the Breuerian constancy was a search for an optimum physiological functioning of the nervous system that involved an available, but not excessive, quantity of positive tonic energy, Freud proposes a principle based on physics. The "inertia" at the root of his system is grounded in a search for a total absence of excitation, or the fastest possible discharge of any quantities of excitation so as to restore a level of zero.

If "constancy" might appear to be a physiological term, Freud's deliberate choice to give the term from physics of "inertia" to this fundamental principle of the psychical function, is not without significance. When associated with the reference to the "general laws of motion," it is explicitly positioned within the epistemological framework of Newtonian dynamics (Freud, 1895). In his *Principia Mathematica*, after having defined the quantity of matter and the quantity of motion, Newton states the three axioms that make up the "laws of motion." The first law, termed law of inertia (Ansermet, 2009), is defined in these terms:

> *"Every body perseveres in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed thereon." (Newton, 1687/1846, p. 83)*

It is relevant to underline that, from a formal point of view, the Freudian argument mimics the structure of Newton's text: definition of the fundamental theory of quantity (the neuronal excitation in motion), followed by a description of the laws or principles that regulate the movement of the quantities – the principle of inertia appearing as the first axiom. Furthermore, the choice of the term "inertia" posits, from an epistemological point of view, the perspective adopted by Freud as explicitly physicalist, through a reference to classical Newtonian dynamics, and that before the slightest references to any strictly organic principle. Finally, if we refer back to the axioms of *Principia Mathematica*, the meaning of the term "inertia" is to be understood in relation to the first law of motion, and more specifically to the first example given: a body at rest stays at rest unless it is acted upon by a force. Thus, the nervous system, if we imagine a mythical primal state that would be the absence of any excitation, should tend to retain that "zero level."

Newton goes on to give a second law of motion, one that specifies the event wherein a force acts upon a body, therein moving it out of its state of inertia. This second law thus allows the definition of the principles that govern the change in quantity of motion of the system, the body no longer finds itself in the ideal situation of inertia, and is subjected to the action of a driving force: "The alteration of motion is ever proportional to the motive force impressed" (Newton, 1687/1846, p. 83). Now, Freud finds he also has to consider a second principle, a "secondary function," that describes the functioning of the neuronal system went it can no longer conserve the zero level. If we apply the first two Newtonian laws of motion to the nervous system we could, as did Freud, imagine that it is initially in a state of repose (the "level = zero"); and that it would tend to remain in that state of repose, until the moment when a given force introduces into it a quantity of movement. However, we can remark, as do Laplanche and Pontalis, that "the relationship between Freud's use of the principle of inertia, and its application in physics, remains quite flexible" (Laplanche and Pontalis, 1967b, p. 340). It remains quite flexible in so far as that in physics inertia essentially consists in a property of bodies in motion, whereas "for Freud, it is not a property of the envisaged motivation, that is to say excitation, but an active tendency of the system within which the quantities move" (Laplanche and Pontalis, 1967b, p. 340). Nevertheless, the principle of inertia does consist in the tendency of the "particles of matter" that are neurons to divest themselves of the quantities of excitation that traverse them, and therefore to return to their state of repose. Finally, the third Newtonian law of motion, according to which "To every action there is always opposed and equal reaction: or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts." (Newton, 1687/1846, p. 83), could offer some similarities with the Freudian definition of the discharge during the reflex movement. The neurons "neutralizing the reception of *Qi* by giving it off." (Freud, 1895, p. 296), through the reflex movement that amounts to a mode of discharge:

> *"A primary nervous system makes use of this Qi which it has thus acquired, by giving it off through a connecting path to the muscular mechanism, and in that way keeps itself free from stimulus. This discharge represents the primary function of the nervous system." (Freud, 1895, p. 296)*

The application of the third law could thus be understood in these terms: faced with the introduction of a quantity of excitation considered as a driving force, the conservation of the initial sate of repose or of non-excitation (the application of the principle

of inertia), can be assimilated to a reaction that, in Newton's terms, would be "always equal and opposite" to the action. The introduction of a quantity of excitation into the system, and its discharge through the reflex, are therefore the result of the action of equal quantities acting "in opposite directions."

If, in the words of Laplanche and Pontalis, the Freudian construct can in some sense "appear to be an aberration from the viewpoint of the life sciences" (Laplanche and Pontalis, 1967a, p. 326), in that it postulates a first principle of the functioning of the nervous system that is "the negation of all stable difference of level" (Laplanche and Pontalis, 1967a, p. 326), it is relevant, in order to dissipate this aporia, to put it back into context within the radically physicalist epistemological model that Freud uses. Freud remains loyal to the tradition of the Hemholtz School, and to contemporary research in thermodynamics. The principle of constancy as it is described by Breuer is explicitly to be situated within the framework of the first law of thermodynamics, that of the conservation of energy. Thus, according to him, the nervous system would endeavor to keep constant an optimum level of tonic energy to ensure its smooth functioning. However, the Freudian principle of inertia, and the death drive that became its logical continuation in the 1920, could be explained by reference to entropy, that is, to the second law of thermodynamics.

FROM THE FIRST LAW OF THERMODYNAMICS ON THE CONSERVATION OF ENERGY, TO THE FORMULATION OF THE SECOND LAW

The principle of the conservationof energy, based on the work done on heat machines by Carnot [described in his 1824 memoir *Reflections on the Motive Power of Heat.* (Carnot, 1824/1897)], made it possible to formulate an equation for the transformation of heat into a quantity of motion. However, this research was based on the model of an idealized machine, whose utopian output would not be subject to any loss. The beginnings of thermodynamics had therefore neglected to take into consideration the fact that what steam engines consume, irreversibly disappears; no heat machine will restitute the coal it devours. The formulating of the second law of thermodynamics thus stems, according to Prigogine and Stengers (1979), from the transition between a formalization of the transformation of energy within a reversible equation, to the reality of the losses that this conversion entails. According to the law of the conservation of energy, the mechanical work produced and the reduction in the difference in temperature are thus connected in an ideal way through a reversible equivalence, in so far as that the same machine, working in reverse, could restore the initial difference. The taking into account of the losses that, for any real engine, result in an output inferior to the ideal output predicted by this equation, signals therefore the advent of a new science. A science that is no longer based on idealization, but on nature itself, including its "losses." In this way, the concept of "irreversibility" makes its appearance in physics: there are *irreversible* disturbances, losses that diminish the output of heat machines, which are linked to a dissipation of energy (Laplanche and Pontalis, 1967a).

In 1852 William Thomson formalized this observation by stating the second law of thermodynamics in his papers on the *Dissipation of Mechanical Energy* (Thomson, 1852, In: Locqueneux, 2009). This law states that, in the course of the production of mechanical work from a heat source, "equal quantities of heat are put out of existence" (Thomson, 1852, In: Locqueneux, 2009). This irreversible dispersal of heat is, in the context of thermodynamic machines, synonymous with a loss of output; something that Thomson presents as a "tendency toward a universal degradation of mechanical energy" (Prigogine and Stengers, 1979, p. 185). According to Prigogine and Stengers, by pronouncing the second law of thermodynamics, Thomson accomplishes "the vertiginous leap from motor technology to cosmology" (Prigogine and Stengers, 1979, p. 186), in so far as that he accomplishes an epistemological revolution that renders the world of Laplace, with its simple conservative and eternal ideal machine, definitively obsolete. Henceforth, the world can be described as a machine within which the conservation of heat in motion could only be achieved at the cost of an irreversible wastage, owing to the dissipation of a given quantity of heat (Prigogine and Stengers, 1979). From this principle follows a new description of the world: "the differences that produce an effect are continuously diminishing within nature" (Prigogine and Stengers, 1979, p. 185), and the world in the course of these conversions of energy "depletes these differences" to finally reach a state of thermal equilibrium where no difference that produces an effect would subsist.

We find here a certain resemblance between the Freudian principle of inertia assimilated to a "negation of all stable level of difference" (Laplanche and Pontalis, 1967a), and the second law of thermodynamics according to which the world tends toward an annihilation of the differences that produce effects, in a search for a state of equilibrium. Furthermore, it should be pointed out that, the second law contributed to giving a new importance to the question of time in physics – whereas the Laplacian world, conceived within its reversible unity, had to some extent, not so much resolved, but pushed aside this question. With the advent of the concept of irreversibility in physics, time also introduces itself into that discipline, in the guise of an evolution toward homogeneity and death (Prigogine and Stengers, 1979). Now, this understanding of a temporal evolution toward a state of homogenous equilibrium, equivalent to death in a living organism (that is, a return to the inanimate), is already in essence in the Freudian definition of the principle of inertia; and will find its clearest formulation with the introduction of the "death drive" in *Beyond the Pleasure Principle* (Freud, 1920).

In 1865, Rudolf Clausius (Clausius, 1865, In: Locqueneux, 2009) produced a mathematical formulation that made it possible to include both the reversible transformations of classical mechanics, as well as the irreversible physicochemical transformation that conserves energy while not being reversible (that is to say when a reversal of the functioning of the system cannot make it return to its initial state, as is the case with friction, where the motion is converted into heat) (Prigogine and Stengers, 1979). Clausius posits a state function *S*, which

he calls "entropy," so as to make a distinction between these two cases. From there Clausius concludes that the principle of conservation of energy, such as Helmholtz had recognized as a general principle, is contradicted by the second law. Thus, if the first law states that:

> *"A form of energy can transform into another form of energy, but the quantity of energy can never be lost; on the contrary, the total energy existing in the universe remains constant, just as the quantity of matter." (Clausius, 1865, In: Locqueneux, 2009, p. 248)*

therein proposing a concept of the universe as a whole, as absolutely irreversible, eternally performing its revolutions; the second law, that is applicable in a general way to all transformations that occur in the universe, reveals that:

> "the transformations do not need to be represented in equal quantities in opposite directions, but the difference can only occur in one determinate directions [. . .]. The outcome of this is that the state of the universe must continuously and increasingly change in one determinate direction." (Clausius, 1865, In: Locqueneux, 2009, p. 248)

THE APPLICATION OF THE PRINCIPLE OF ENTROPY TO THE WHOLE FORMED BY THE ORGANISM AND ITS ENVIRONMENT: THE THERMODYNAMIC ORIGIN OF THE FREUDIAN CONCEPT OF "DEATH DRIVE?"

Mechanical work tends increasingly to turn into heat, there is therefore an increasing and irreversible dissipation of heat since: "heat, that constantly passes from the warmer bodies to the cooler bodies, consequently rendering the temperature equal on both sides, will gradually be distributed in an increasingly equal way; a determinate equilibrium will be established between the heat emanating from the ether, and the heat that is in the bodies" (Clausius, 1865, In: Locqueneux, 2009, p. 248). Clausius, therein, introduces the theory of a general tendency toward a state of equilibrium. A tendency where the transformations will gradually come to end in a stable state, without variations in levels, and where no further difference resulting in an effect could take place. The tendency, according to the Freudian principle of neuronal inertia, for neurons to discharge, can be assimilated to a search for a state of equilibrium, the "level = 0."

Based on his observation of heat machines, Clausius sought to formulate as a law this progressive, diachronic, change toward a state of equilibrium, that is defined as "the state toward which the universe gradually tends." Thus, Clausius also makes the "the vertiginous leap from motor technology to cosmology" (Prigogine and Stengers, 1979, p. 186). In this law, he gives the name "entropy" to the vastness that represents "The sum of all the transformation that must occur to bring a body or a system of bodies to its current state" (Clausius, 1865, In: Locqueneux, 2009, p. 248), and concludes from this that "in all natural phenomena, the total value of entropy can only increase without ever decreasing" (Clausius, 1865, In: Locqueneux, 2009, p. 248). He sums up this change, that constantly takes place everywhere in nature, with the following law, that has remained famous:

> *"the entropy of the universe tends to a maximum." (Clausius, 1865, In: Locqueneux, 2009, p. 248)*

Consequently, according to Prigogine and Stengers, Clausius introduces hereby an "arrow of time" into physics, in that entropy can only increase in the course of time or remain constant. The increase in entropy therefore translates into an irreversible temporal evolution of the system, an evolution of a spontaneous kind. Thus, for every isolated system, the future could be defined in physics as the direction in which entropy increases. The second law implies therefore that for a given isolated system, not all evolutions are of equal value: equilibrium would appear to be a veritable "attractor" for states of non-equilibrium. The irreversible increase of entropy describes a nearing of a system to a state that attracts it, that it "prefers," and from which it no longer spontaneously distances itself; therefore, a nearing that is irreversible (Prigogine and Stengers, 1979). However, the second law does not invalidate the first law of conservation of energy; on the contrary it encompasses it in a generalized theory. Reversible changes would thus be extreme cases, in which nature has as much propensity for the initial state as for the final one (Planck, 1941).

If the universe's entropy "tends toward its maximum" then, according to Clausius, the more the universe draws close to that limit state, the more the opportunities for new changes disappear. When that state is reached, no further change would occur, and the universe finds itself in a "persistent state of death" (Clausius, 1865, In: Locqueneux, 2009, p. 249). These considerations make it possible to reread Freud's hypothesis of the death drive in the light of the physics model, which had from the outset been the epistemological paradigm for his initial theoretical thinking. If we apply to living organisms this tendency of the universe toward a state of equilibrium, defined by the irreversible absence of all discernible motion and all difference in tendency – in other words equivalent to a definitive death – we can envisage a closed system consisting of the unit "organism-environment." The second law implies that within this system the different levels of energy tend toward equaling out, in such a way that the final state would be a state of equilibrium. The state toward which the system would tend would therefore be equivalent to "the reduction of the organism's internal energy that returns it to the inorganic state" (Laplanche and Pontalis, 1967a, p. 326). Now, let us recall that in 1920, Freud describes the conservative or "regressive" character of the death drives as originating in "the coming to life of inorganic substance" (Freud, 1920, p. 44) and that it "seek(s) to restore the inanimate state" (Freud, 1920, p. 44). If the definition of the death drive as the tendency of the living organism to return to an inorganic state can, in the first instance, "seem an aberration from the point of view of the life sciences – in so far as that it seeks to infer an organism with its vital aptitudes, its adaptative functions, its energy levels, from a principle that is the negation of all constant level of difference" (Laplanche and Pontalis, 1967a, p. 328) – it appears much more coherent within a physics model. Furthermore, it should be pointed out that the first ambition of the Helmholtz School, to which Freud was heir, had as specific

objective the application of the physical laws of thermodynamics to the study of living organisms.

The thermodynamics that inspired Freud described how systems reached an equilibrium characterized by a maximum of entropy. In contrast[1], as Prigogine (1978, 1981) pointed out, when systems are far from equilibrium and driven by a large energy current, entropy may decrease and ordered patterns form. Current research focuses precisely on the self-organization of systems far from equilibrium. The neuronal dynamics, with the discharges that Freud envisaged, find echo in the theory of self-organized criticality (Bak, 1996; Vespignani and Zapperi, 1998). Major advances were achieved by adapting the evolution equations of statistical physics (Fokker–Planck equations) to out-of-equilibrium, open systems (Seifert, 2005; Tomé, 2006; Esposito et al., 2009; Jeffery et al., 2019). It is quite remarkable, in view of Freud's affinity for Helmholtz work, that a thermodynamic potential, the free energy, appears to be the quantity that best describes the steady state of a system driven out of equilibrium because of its strong interactions with its environment (Jarzynski, 1997; Evans and Searles, 2002; Friston, 2019). Based on these new ideas, the "death drive" might be recast as a natural tendency of certain out-of-equilibrium systems to reach a steady-state characterized by a minimization of free energy. Indeed, there have been attempts to connect Freudian notions of free (unbound) energy to the variational free energy that figures in theoretical neurobiology and statistical mechanics (Carhart-Harris and Friston, 2010). If this is possible, it would mean that the death drive might now be cast in a way that is formally similar to the way Newtonian mechanics was recast in a "principle of least action." This would constitute a major advance in Freud's "*Project for a Scientific Psychology*."

CONCLUSION: THE DEATH DRIVE: BEYOND AN ANTITHESIS BETWEEN A PHYSICAL, OR BIOLOGICAL, PARADIGM

If we reconsider this idea in the light of the physiological tradition, we can point out that the Freudian death drive is not in complete opposition to the thinking of the French School. The experimental research of Claude Bernard had contributed to focusing biology away from the vitalist concept according to which (as was argued by Bichat amongst others) life should be defined as "the sum of the forces that resist death" (Bichat, 1852, p. 1). Alongside the discovery of the second law of thermodynamics, Bernardian physiology had overthrown this definition in favor of a concept of death as an integral part of the vital phenomena. Something that Bernard encapsulates in the twofold aphorism: "life is death" and "life is creation" (Bernard, 1885, p. 40), in which the two terms are indissociable and form a dialectical whole (Prochiantz, 1990). Bernard's research on the physiology of respiration, nutrition, and organic combustion, brought to light that the destruction of tissues is the consequence of these vital functions. Thus, he rendered null all the vitalist physiology that rested on the opposition between a vital force, and a natural, physicochemical, tendency to move toward death (Prochiantz, 1990). According to Bernard, science had thus put an end to the split between two kinds of property within the living organism, the physical properties and the vital ones. Properties that were understood as being in a constant state of opposition and strife. So, no "grip" held by the vital properties within the organism, in so far as that the vital functions are regulated by a strict physicochemical determinism. Now, this critique of Bichat's definition according to which life constitutes the sum of the forces that are in opposition to death, could, according to Prochiantz, also be interpreted as an argument against any concept of life as a singular point of resistance to the second law of thermodynamics. That is to say, as a structure that, at a given point, opposes increasing entropy (Prochiantz, 1990). Life then, would be destruction itself, compensated at each moment by the process of creation. In this respect, life could no longer be defined as that which resists destruction, or the increasing entropy of the universe if we adopt the terminology of physics. If we take into account this evolution of biology, made possible notably by the Bernardian revolution in physiology, the ideas formulated by Freud throughout his work (from the principle of inertia in 1895 up to the death drive in 1920) can, although linked to the physicalist epistemological framework, also resonate with this change in the biological understanding of living organisms. In view of these considerations they would no longer appear as an "aberration" from the view point of the life sciences, but on the contrary would revisit some of the questions raised in biology at the end of the 19th century.

AUTHOR CONTRIBUTIONS

JT is the main contributor of this manuscript as part of her Ph.D. thesis. J-PA, FA, and PM as supervisors, contributed to the conception and development of the research, and critically revised the manuscript for intellectual content.

[1] We thank the referee for suggesting to bring our historical analysis in the light of current research, part of which is the topic of the present issue.

REFERENCES

Ansermet, J.-P. (2009). *Mécanique, Volume I*. Lausanne: PPUR presse polytechniques.

Ansermet, J.-P. (2019). "Quantité-qualité," in *Proceedings of the Séminaire de la fondation Agalma*, (Geneva: Fondation Agalma).

Assoun, P. L. (1981). *Introduction à L'épistémologie Freudienne*. Paris: Payot.

Bak, P. (1996). *How Nature works: The Science of Self-Organized Criticality*. New York, NY: Springer Verlag.

Bernard, C. (1885). *Leçons Sur Les Phénomènes Communs Aux Végétaux at Aux Animaux*. Paris: Baillère & Fils.

Bichat, X. (1852). *Recherches Physiologiques Sur la Vie et la Mort*. Paris: Masson.

Carhart-Harris, R. L., and Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain* 133, 1265–1283. doi: 10.1093/brain/awq010

Carnot, S. (1824/1897). *Reflections on the Motive Power of Heat*. New York, NY: John Wiley & Sons.

Clausius, R. (1865). Le second principe de la théorie mécanique de la chaleur. *C.R.* 60, 1025–1027.

Esposito, M., Harbola, U., and Mukamel, S. (2009). Nonequilibrium fluctuations, fluctuation theorems, and counting statistics in quantum systems. *Rev. Mod. Phys*. 81, 1665–1702.

Evans, D. J., and Searles, D. J. (2002). The fluctuation theorem. *Adv. Phys.* 51, 1529–1585.

Freud, S. (1895). *A Project for a Scientific Psychology, S.E., 1* (London: Hogarth), 283–397.

Freud, S. (1900). *The Interpretation of Dreams, S.E., 4–5*. London: Hogarth.

Freud, S. (1920). *Beyond the Pleasure Principle, S.E., 18*. London: Hogarth.

Freud, S. (1925). *An Autobiographical Study, S.E., 20* (London: Hogarth), 3–70.

Freud, S. (2017). La structure des éléments du système nerveux. *Essaim* 38, 119–130.

Freud, S., and Breuer, J. (1895). *Studies on Hysteria. S.E., 2*. London: Hogarth.

Friston, K. (2019). A free energy principle for a particular physics. *arXiv* [Preprint]. Available at: https://arxiv.org/ftp/arxiv/papers/1906/1906.10184.pdf (accessed June 24, 2019).

Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett*. 78, 2690–2693.

Jeffery, K., Pollack, R., and Rovelli, C. (2019). On the statistical mechanics of life: schrödinger revisited. *Entropy* 2019:1211.

Jones, E. (1953). *The Life and Works of Sigmund Freud*, Vol. I. New York, NY: Basic Books.

Laplanche, J., and Pontalis, J.-B. (1967a). "Principe de constance," in *Vocabulaire de la psychanalyse* (Paris: PUF).

Laplanche, J., and Pontalis, J.-B. (1967b). "Principe d'inertie neuronique," in *Vocabulaire de la psychanalyse* (Paris: PUF).

Locqueneux, R. (2009). *Histoire de la Thermodynamique Classique : de Sadi Carnot à Gibbs*. Paris: Belin.

Newton, I. (1687/1846). *The Mathematical Principles of Natural Philosophy. A. Motte (Trans.)*. New York, NY: Daniel Adee.

Planck, M. (1941). *Initiation à la Physique*. Paris: Flammarion.

Prigogine, I. (1978). Time, structure, and fluctuations. *Science* 201, 777–785.

Prigogine, I. (1981). *From Being to Becoming, Time and Complexity in the Physical Sciences*. New York, NY: W. H. Freeman.

Prigogine, I., and Stengers, I. (1979). *La Nouvelle Alliance*. Paris: Gallimard, 188–189.

Prochiantz, A. (1990). *Claude Bernard, la Révolution Physiologique*. Paris: PUF.

Seifert, U. (2005). Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Phys. Rev. Lett*. 95:040602.

Thomson, W. (1852). Deux mémoires sur la théorie dynamique de la chaleur. *J. Math. Pures Appl.* 17, 209–252.

Tomé, T. (2006). Entropy production in nonequilibrium systems described by a fokker-planck equation. *Braz. J. Phys.* 36, 1285–1289.

Tran The, J., Magistretti, P., and Ansermet, F. (2018). The epistemological foundations of freud's energetics model. *Front. Psychol.* 9:1861. doi: 10.3389/fpsyg.2018.01861

Vespignani, A., and Zapperi, S. (1998). How self-organized criticality works: a unified mean-field picture. *Phys. Rev. E* 57, 6345–6362.

Conflict of Interest: 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 Psychology

CONCEPTUAL ANALYSIS
published: 17 March 2020
doi: 10.3389/fpsyg.2020.00366

Entropy, Free Energy, and Symbolization: Free Association at the Intersection of Psychoanalysis and Neuroscience

Thomas Rabeyron[1,2]* *and Claudie Massicotte*[3]

[1] Interpsy, Université de Lorraine, Nancy, France, [2] University of Edinburgh, Edinburgh, United Kingdom, [3] Young Harris College, Young Harris, GA, United States

Both a method of therapy and an exploration of psychic reality, free association is a fundamental element of psychoanalytical practices that refers to the way a patient is asked to describe what comes spontaneously to mind in the therapeutic setting. This paper examines the role of free association from the point of view of psychoanalysis and neuroscience in order to improve our understanding of therapeutic effects induced by psychoanalytic therapies and psychoanalysis. In this regard, we first propose a global overview of the historical origins of the concept of free association in psychoanalysis and examine how Freud established its principles. Then, from Freud's distinction between primary and secondary processes, we proceed to compare the psychoanalytic model with research originating from cognitive psychology and neuroscience. The notions of entropy and free energy appear particularly relevant at the intersection of these different domains. Finally, we propose the notion of symbolizing transmodality to describe certain specificities of symbolization processes within free association and we summarize the main functions of free association in psychoanalytic practices.

Keywords: free association, psychoanalysis, symbolization, neuropsychoanalysis, free energy, entropy, primary processes

OPEN ACCESS

Edited by:
Christoph Mathys, International School for Advanced Studies (SISSA), Italy

Reviewed by:
Patrick Connolly, Hong Kong Shue Yan University, Hong Kong
Daniela Flores Mosri, Universidad Intercontinental, Mexico

***Correspondence:**
Thomas Rabeyron
thomas.rabeyron@gmail.com

Specialty section:
This article was submitted to Psychoanalysis and Neuropsychoanalysis, a section of the journal Frontiers in Psychology

Received: *24 October 2019*
Accepted: *17 February 2020*
Published: *17 March 2020*

Citation:
Rabeyron T and Massicotte C (2020) Entropy, Free Energy, and Symbolization: Free Association at the Intersection of Psychoanalysis and Neuroscience. Front. Psychol. 11:366. doi: 10.3389/fpsyg.2020.00366

INTRODUCTION: FREE ASSOCIATION AS THE CORNERSTONE OF PSYCHOANALYTIC PRACTICES

The effectiveness of psychoanalysis and psychodynamic approaches in the treatment of mental disorders has been the object of numerous empirical studies (Shedler, 2010; Steinert et al., 2017). Current work aims to understand the way such approaches operate, what distinguishes them from other therapeutic methodologies, and their efficacy for long-term psychic transformations (Leuzinger-Bohleber et al., 2019; Woll and Schönbrodt, 2019). Free association – presented by Freud (1913) as the "fundamental technical rule" of psychoanalysis – is often considered as the cornerstone of psychoanalytic practices (Bollas, 2008). Barratt (2016, 2017) thus reminds that "Freud continued to assert consistently that the method of free association is the *sine qua non* of his discipline" (2017, p. 39) and proposes a return to the discipline's roots relying on free associative praxis. Similarly, for Scarfone (2018), "Free association is really a most distinctive and foundational part of the procedure we call psychoanalysis" (p. 468). Free association thereby appears to be a key concept to examine the modalities and effects of psychic transformation proceeding from

psychoanalysis and psychodynamic therapies. In these settings, free association defines the way the patient may spontaneously and unreservedly say anything that comes to mind. The clinician will then be attentive to the way in which the patient goes from one representation to another with more or less fluidity during the therapeutic sessions.

Following these previous lines of research, this article proposes a synthesis concerning the fundamental value of free association processes during psychoanalysis and psychoanalytic psychotherapies. It first presents the historical origins of the concept of free association in psychoanalytic theory, then discusses its development within research in cognitive psychology (Kahneman, 2003, 2011), neuroscience (Friston, 2009; Carhart-Harris and Friston, 2010), and neuropsychoanalysis (Solms and Turnbull, 2011). Despite significant distinctions within these models, we focus on the connections between psychoanalytic and neuroscientific concepts to highlight the heterogeneity of psychic modes of symbolization (Roussillon, 2015), thus developing earlier observations in the fields (Mancia, 2006; De Masi et al., 2015). In this regard, we will underline, as proposed by Cieri and Esposito (2019), how "free association offers a clear and sharp path with cognitive science, free energy neuroscience, and computational psychiatry in order to create a consistent and solid connection between the psychological and neuroscientific views" (p. 5). Free association will thus emerge as a particularly fruitful concept to specify the understanding of therapeutic models through a dialogue between psychoanalysis and neuroscience (Magistretti and Ansermet, 2010; Panksepp and Solms, 2011; Yovell et al., 2015; Rabeyron, 2016).

ORIGINS OF FREE ASSOCIATION IN PSYCHOANALYTIC THERAPIES AND PRACTICES

Historically, reflections on the activity of thought, and the free association which characterizes it, emerged during the 18th century through the "exteriorized" conceptions of Franz-Anton Mesmer. His notion of "animal magnetism" as a "universal flux" that must be harmoniously reordered through various processes (magnetism, passes of hands, etc.) offered a view of mental energy as an external force (Méheust, 1999). This first attempt to represent a "psychic flux" gradually became more "internalized" with the development of psychoanalysis (Laplanche, 1987; Roussillon, 1992). Yet, despite the evolution of psychological theories since Mesmer, the idea of a "flux" that could become "blocked," thus giving rise to various forms of psychopathology, never completely disappeared, and vestiges of such ideas can still be found in present theories of free association (Roussillon, 2009, 2012; Donnet, 2012).

During the 19th century, Pierre Janet evoked "points of fixation" in psychic activity to describe such obstruction, and Freud (1895) pursued this idea in his *Project for a Scientific Psychology*, yet added the hypothesis that specific "primary defenses" led to these points of fixation. Freud's originality also consisted in his conception of these defense mechanisms being the consequence of traumas and previous life experiences related to the subject's affective and sexual life. He explained that an "inhibiting lateral investment" could protect the subject from previous traumatic events by inducing a blockage of free association. This defensive architecture would then limit the patient's associative capacities[1]. Later, Freud further remarked that these fixations originated from a kernel "of historical truth" (Freud, 1937) – for example, a traumatic experience – which would reemerge through the repetition compulsion because of a "weakness of the power of synthesis" of the ego (Freud, 1941, p. 229).

Freud then supposed that mental functioning and psychopathology could be studied, thanks to free association, according to the particularities of the associative flow and that patients could work through these fixations via free association. He began to use this process with hypnosis and was asking his patients the first words that came to mind while he placed his hand on their forehead. He then conceptualized free association without hypnosis during his work with Emmy Von N. (Freud and Breuer, 1895) and specified his ideas in *The Interpretation of Dreams* (Freud, 1900). Freud showed that the latent content of the dream could be deciphered through the thoughts the patient spontaneously associated with the dream. Freud later used the same technique in *Psychopathology* of Everyday Life (Freud, 1901) to understand slips of the tongue, forgotten words, etc. Then, he employed free association with Freud (1905) to analyze several of her symptoms and again with the Freud (1909) in order to understand the source of the latter's obsessional behaviors. In his essay *On Beginning the Treatment*, Freud (1913) proposed a clear metaphor to describe the mechanisms of free association to his patients: "Act as though, for instance, you were a traveler sitting next to the window of a railway carriage and describing to someone inside the carriage the changing views which you see outside" (1913, p. 135). For Freud, this method of investigation of psychic reality, and its unconscious processes, also served a therapeutic function and could help the patient release the flow of the activity of thought. During the psychoanalytic treatment, Freud would help the patient to deploy free associations in order to restore or catalyze "blocked" psychological processes and conflicts. Freud's works about free association thus defined the way in which one passes spontaneously from one idea to another in the psychoanalytical setting and the connections between free association, psychic functioning, psychopathological disorders, and the therapeutic effects of the psychoanalytic treatment.

Thus, it was largely through the free association method that Freud came to analyze the different layers of the psyche and to distinguish between primary and secondary processes corresponding to different "treatments" of psychic energy (Freud, 1915). In the Freudian model, primary processes characterize the unconscious system, while secondary processes are associated with the preconscious-conscious system. In primary processes, psychic energy is said to flow more "freely" and to shape thing-(re)presentations according to the hallucinatory satisfaction of desire. The dream emerges here as a prototype of this

[1]For an overall examination of Freud's approach to the binding processes and their relationship toward the limitation of free association processes, see Holt (1962), who underlines how "Freud used binding (and its opposites, freedom or mobility of cathexis) in over a dozen different ways as his theory developed" (p. 522).

type of primary functioning in which operate deformation mechanisms of great malleability, such as displacement and condensation. For Freud, a model of "identity of perception" prevails in primary processes as the psyche appears to reproduce through hallucinations previous pleasant sensorial and perceptive experiences. Within secondary processes, on the other hand, psychic energy has to be bound for the word-(re)presentations to be more stable. The mode of satisfaction appears to become secondarized and "identity of thought" now prevails, for the source of pleasure in secondary processes is no longer the identical reproduction of a previous pleasurable experience but the symbolic thinking associated with the initial pleasant experience. In its relation to the world, the psyche has thus sacrificed part of its freedom in its relation to pleasure in order to adapt to reality, and the associative flux thereby finds itself diminished.

Freud supposed that the analyst should be in a specific state of mind called "free floating attention" while the patient is free associating. In this way, analysts might use their own unconscious to decipher the unconscious of the patient. Contemporary psychoanalytical models of free association have since insisted on this aspect and claim that free association is only fully effective when coupled with this form of free association coming from the analyst. This "shared" free association, or co-associativity[2] (Roussillon, 2011) implies that the patient associates freely in the presence of the clinician and addresses oneself through the other. Alterity thus emerges as a fundamental dimension of the associative process: one may not freely express the secrets of one's most intimate psychic life in a solipsistic way; rather, one must find the conditions to deploy free association in the intersubjective relationship (Barratt, 2017).

Various psychoanalysts since Freud have argued that this shared free association operates at a very "primary" level through a form of "co-thinking" (Widlöcher, 1996, 2010) or "co-psycheity" (Georgieff, 2010) particular to psychodynamic psychotherapies and the psychoanalytic setting. Thanks to the transference process, the spontaneous free associations of the analyst may reflect some unelaborated aspects of unconscious processes in the patient's own associativity. We are thus dealing with an "analytical third," that is to say a melting of the free association processes of the patient and analyst at a very primary level (Ogden, 1994). Eschel (2006) describes more precisely a process of "twogetherness" constituting a form of "associativity of presence" when the relation to the psychoanalyst is established primarily through affects. This shared and primary associativity becomes the breeding ground necessary for the emergence of a "moment of meeting" (Stern, 2004) during which both clinician and patient feel that a step has been made toward maturation and symbolization processes.

Some yet unmetabolized experiences will then "blister" (*boursoufler*) the patient's free association and behaviors in the psychoanalytic setting in order to be shared and recognized (Roussillon, 2012; Lothane, 2018). The patient may act out – the Freudian *agieren* – what remains unelaborated from previous sufferings and pathological relationships. For example, this process may give rise to the "fear of breakdown" described by Winnicott (1963), a fear which re-emerges in consequence of early primitive sufferings. It may also occasion the return of traumatic experiences in hallucinatory forms during the therapeutic sessions (Botella and Botella, 1990). These past traumatic experiences will leave "knots" or "marks" on free association, the latter being "directed at unraveling the knots in the patient's psyche" (Scarfone, 2018, p. 474). The work of integration and transformation operating through the "unbridled" free association in the clinical setting therefore requires that the unelaborated experience be expressed, notably through the transfer, "fragment by fragment," or "piece by piece" as suggested by Freud (1913). A "transfer" and a shared associativity then allow a translation process of the past traumatic experiences. This process permits the patient to "re-feel" or "re-know" an experience that has remained unmetabolized in order to improve reflexive awareness, which is catalyzed and condensed by the clinical setting. Free association and reflexivity thus share a need to deploy themselves through exteriority: what cannot be represented and symbolized through intrapsychic processes must be "externalized" thanks to free association and the intersubjective relationship in order to be elaborated.

FREE ASSOCIATION AND FREE ENERGY FROM THE POINT OF VIEW OF COGNITIVE PSYCHOLOGY AND NEUROSCIENCE

To what extent are these models of free association developed in psychoanalysis in line with recent work in the field of cognitive psychology and neuroscience? A first comparison emerges through the work of Kahneman (2003) who focuses on the understanding and modeling of reasoning biases, studying them through various ingenious experiments. Kahneman (2003, 2011) proposes a division of consciousness according to two principal modes of thinking. He calls the first "System 1" to describe reasoning fallacies emerging from a fast and imprecise activity of thought linked to intuitive functioning[3]. This System encompasses automatic feelings and inclinations, is almost instinctive, and yet is shaped by experience. System 1 builds logical causalities outside the sphere of conscious awareness and is easily influenced by phenomena of suggestion and

[2]We will use in this paper the term "associativity" which is a translation of the French term "associativité" used in particular by Roussillon (2011). Associativity is a more general term than free association in the sense that it supposes that the free association process is not reducible to the verbal free association. The latter can take different forms according to various clinical devices: it can be, for instance, focal when it is centered on a dream or a projective test; it can involve several people when it arises from the "associative chain of a group;" and it can be "projected" externally on an object (for example, a painting) during artistic creation used during therapy (Rabeyron, 2017) leading to what Brun (2014) named "associativity of the shapes."

[3]For instance: a baseball bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much is the ball? Even among students from the Ivy Leagues, one out of two individuals responded erroneously (the correct answer is $0.05).

priming[4]. Mood and cognitive engagement also have a major impact on this System's functioning. System 1 is sensitive to the "halo effects" and produces a set of approximations in reasoning. Kahneman concludes that the human brain naturally favors the slightest effort and prefers to stick to the most accessible information. When the approximations thus produced are not secondarized – that is, validated by System 2 –, subjects tend to make more cognitive mistakes. Kahneman describes System 1 as an "associative machine" functioning through logics of "associative coherence," in the sense that it spontaneously and automatically constructs meaning from underlying causal links.

Through an original methodology, Kahneman's research echoes Freud's attempt at mapping psychic heterogeneity through the distinction between primary and secondary processes. Kahneman and Freud's approaches may be compared through the following table inspired from Roussillon (2001) and Kahneman (2003).

System 1 – Primary processes	**System 2 – Secondary processes**
Quick temporality	Slow temporality
Automatic	Reflexive
Unconscious	Conscious
No negation	Negation
Intuitive	Rational
Perceptive	Conceptual
Pleasure principle	Reality principle
Free energy	Bound energy
High entropy	Low entropy

Although these two models do not overlap entirely, it is interesting that, despite very different methodologies, both Freud and Kahneman find two main "layers" of psychological functioning whose characteristics can be translated from one model to another. We could consider the S1 and S2 described by Kahneman as the expression of primary and secondary processes at a cognitive level of functioning even if distinctions remain: Kahneman analyzes psychic modes of functioning primarily in terms of cognitive and reasoning mechanisms, while Freud presents a theory of the psyche that deals primarily with its psycho-affective construction. One might also add that Freud is asking the question of "why," while Kahneman focuses on "how" the psyche functions through these two processes. Yet, Freud and Kahneman's theories converge through their understanding of the fundamental bi-polarity of psychic processes which are often working in concert and which leave their "mark" on mental functioning and free association.

We will now turn to the work of Friston (2009) on the free energy principle (FEP) to describe in more detail a second parallel between Freud's work on free association and recent research in cognitive neurosciences, knowing that "in the last 10 years, the FEP has become the royal road in the dialogue between neuroscience and psychoanalysis, *the bridge* between mind and brain" (Cieri and Esposito, 2019, p. 3). Freud initially studied the heterogeneity of psychic functioning according to the way in which the psyche needs to bind and connect nervous energy after sensorial stimulation from the environment. In their *Studies in Hysteria (1895),* Freud and Breuer built upon the theories of contemporary physicists – especially Hermann von Helmholtz – to formulate the distinction between "static" and "kinetic energy," and Freud developed this opposition through the notions of "free energy" and "bound energy" differentiating the primary and secondary modes of psychic functioning. Later, Freud (1920) supposed that "the primary function of the psychic apparatus was to bind the amount of excitation reaching it" and he conceived neurosis as the consequence of a "surprise" taking the form of a fright induced by traumatic events.

These hypotheses join the recent theories of Karl Friston (2013) who reminds that every living organism must resist the second law of thermodynamics, the spontaneous tendency of any physical system to move toward a state of disorganization, that can be measured through degrees of entropy. Friston (2013) supposes that biological organisms must protect themselves against high degrees of entropy which could result in their death. A high entropy level signals a greater level of disorganization and can come from an external source (for instance, from the environment) or from the organism itself (notably through the natural and spontaneous tendency toward disorganization coming from physical and biological properties of matter)[5]. Also drawing upon Helmholtz's hypotheses, Friston suggests that the brain obeys the same principles and constantly produces coherent and predictive representations of the external world in order to limit entropy and its own disorganized states. To limit increases of internal disorganization, the brain develops a Bayesian[6] probabilistic model to determine potential causes of sensations according to prior beliefs and experiences. But this work of prediction is not perfect and sometimes results in a discrepancy between perceptual data from the environment and mental representations supported by the neural network. Friston (2009) calls this discrepancy or disorganization "free energy." It will induce a subjective feeling of surprise, and since the psyche cannot simulate all possible encounters with the environment, states of "surprise" sometimes arise as the consequence of free energy.

[4]Some experiments led by Kahneman confirm the influence of subtle details on the activity of thought, in particular within System 1, and the modes of associativity. For instance, different images placed in front of a donation box will impact the amount of money perceived. Such experiments demonstrate that priming can have an impact on the associativity of thought that is automatic and unconscious.

[5]Actually, entropy could be the fundamental principle that Freud (1920) was looking for in order to describe some inherent properties of living organisms. Entropy produces at the same time energy and excitation (pleasure drive) but can also induce destructivity and, finally, death (death drive) if not contained enough by the organism. This is an important topic from a conceptual point of view that would deserve further development.

[6]Originating from Thomas Bayes, Bayesian statistics consist in deducing probabilities in response to past events. This model now appears more and more frequently within the field of empirical psychology (see Dienes, 2011).

The brain thus constantly responds to interactions with the environment in an enactive way (Ramstead et al., 2019)[7] and these interactions lead to the development of a "generative model" that allows one to make predictions about the environment. The more reliable these predictions are – or the more the brain limits the gap between the internal world and the external world –the lower is the entropy generated in the brain[8] and the fewer are the effects of surprise. Friston also posits that this generative model is organized with a hierarchical structure where higher levels of cerebral functioning exert constraints on the lower levels. Thus, "suppressing free-energy means that each level tries to explain away prediction errors at its own level and in the level below" (2009, p. 295). Friston also describes the complex relationships between these hierarchical levels, and the top–down and bottom–up processes that modify neuromodulation and the mechanisms of "associative plasticity" at a biological and synaptic level (2009, p. 300). His theory, framed by the computational model, therefore offers an understanding of the neuronal constraints of associativity depending on the FEP.

Pursuing Friston's work on free energy and systems theory, Connolly and van Deventer (2017) explain that the FEP operates at various levels of the organism, what these authors refer to as the "scale free principle." They also draw on Hobson et al. (2014) claim that there is a "hierarchical nature of generative or virtual reality models" (p. 11) to suggest that "the predictive model is organized at multiple nested layers, all of which are influenced by the FEP through this recursive feedback process" (p. 11). But "while psychoanalytic mental processes are fundamentally subject to the FEP, they nonetheless also add their own principles of process over and above that of the FEP" (Connolly and van Deventer, 2017, p. 2). The same authors continue: "the level above (which is psychological) cannot violate the FEP. However [. . .] new organizational principles emerge at this level, so that it is not fully explained by the FEP" (p. 7). Consequently, one cannot understand the highest hierarchical levels solely through the FEP because of the emerging properties of the highest hierarchical level of brain functioning. The organization of these higher-order levels will then affect the lower levels from which they originate and influence the activities at lower levels. Thus, subjectivity and free association appear as a functional flow emanating from the neurological system but developing emergent properties impacting in return the underlying biological systems (see **Figure 1**). There are therefore multiple levels in the generative model (sensory systems, memory, self-representation, etc.) and each obeys various operating logics according to an increasing degree of complexification[9]. These levels, which affect each other through recursive loops,[10] communicate and are distinguished by "the existence of a Markov blanket[11] within the brain [that] affords the opportunity for higher levels in the brain to make inferences about lower levels" (p. 11). How, then, might we study these different levels and which are the most fundamental?

Connolly and van Deventer (2017) offer an interesting response: "it is neither possible nor even desirable to build a complete picture of every possible level of organic and neural organization superordinate to the basic level of biological organization which is the FEP, up to the level of interest which is here psychoanalysis. Rather, it is desirable to identify some of the most significant forms of organization that are foundational to psychoanalysis, but superordinate to the FEP, which can build an intelligible bridge between the two" (p. 12). The authors continue: "What would be needed would be a description of the most relevant and proximal layers that most closely influence the level of interest which is that of psychoanalytic regulatory principles" (p. 13). This is exactly what Freud tried to do by showing the main principles organizing psychic reality (pleasure principle, reality principle, principle of constancy, etc.). Likewise, the distinction between primary and secondary processes appears to make up the two most significant levels of mental functioning associated with specific principles, as suggested by both Freud (1900) and Kahneman (2011). As we shall now explore, Solms, Friston, and Carhart-Harris also propose a model that reflects and enriches psychoanalytic models and the modelization of these principles, particularly as they relate to free association.

CONSCIOUSNESS, FREE ASSOCIATION, AND THE DEFAULT MODE NETWORK

Mark Solms has opened an important dialogue between research in contemporary neuroscience – especially the work of Karl Friston – and psychoanalytic models concerning the notion of free energy. In an article entitled "The Conscious Id," Solms (2013), following the work of Panksepp (1998, 2010)[12], criticized

[7]From this point of view, The Free Energy Principle also appears as a way to formalize the notion of autopoiesis and the relationship between the organism and the environment (Friston, 2013).

[8]The relationship with the external world can also be shaped by the actions of the biological organism, actions that can lead to variations in the potential effects of entropy (Friston, 2009).

[9]The ego thus emerges as "an associative structure occupying the higher level of organization of the generative model, that comes to influence lower levels of the hierarchy" (Connolly and van Deventer, 2017, p. 14).

[10]This model appears similar to the work of Aulagnier (1975) distinguishing between originary, primary, and secondary levels to describe the constant work of psychic metabolization underlying representational activity.

[11]On the Markov blanket, Cieri and Esposito (2019) note that "self-organizing systems must have a specific identifiable boundary condition: the so-called Markov blanket, which acts as a protective screen, [. . .] as a veil through which we are able to recognize and distinguish an internal side from an external environment of an organism, inferring the external or internal causes of sensations, perceptions, or changes" (p. 4). As Mellor (2018) suggests, Markov blankets seem to fall under what psychoanalysts have named psychic envelopes, a notion which appears notably in the work of Anzieu (1974) following Bion (1965) writings on the distinction between the container and the contained and the idea of a "membrane" separating conscious and unconscious processes.

[12]Panksepp (2010) describes three levels of control in the brain's emotion-affective processing: primary process (basic-primordial affect), secondary process emotions (learning processes associated with the basal ganglia) and tertiary process (corresponding to neocortical awareness functions). These three different levels are associated with three different levels of consciousness (Solms and Panksepp, 2012): anoetic, noetic, and autonoetic. Consciousness and its relation to emotions are thus considered as the consequence of hierarchic models in which anoetic would correspond to primary processes while noetic and autonoetic would rather correspond to secondary processes in Freud's model

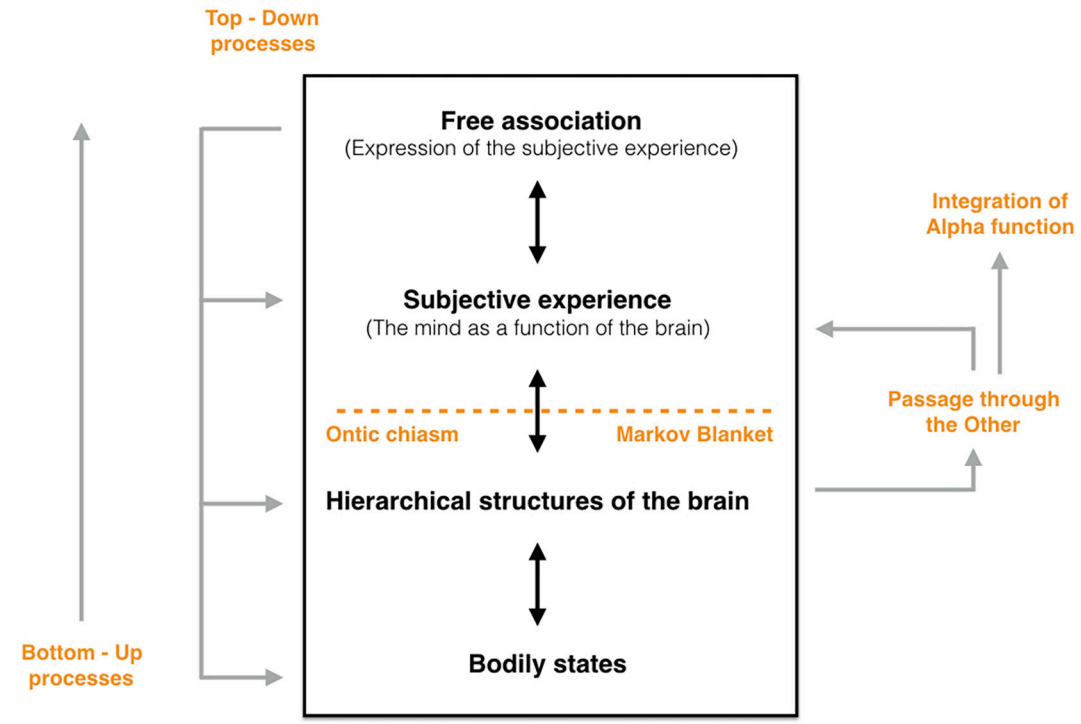

FIGURE 1 | This diagram proposes a synthesis of the passage from the (1) bodily states to (2) the hierarchical structures of the brain to (3) the subjective experience and to (4) the free association that expresses this experience. There exists between the purely biological level and the subjective experience an ontic chiasm which specifies the differences between the processes of conscious psychic functioning. A Markov blanket operates as a space of delimitation and communication between these two levels in the sense that this blanket forms an organizational boundary which allows for the emergence of the subjective experience. This mental functioning emerges through symbolization processes shaped by early intersubjectivity and the passage through the other that characterize it. This intersubjectivity at the origin of thought is gradually internalized and takes the form of what Bion (1965) calls the Alpha function. The highest levels of psychic functioning also influence the lower levels thanks to the top–down processes while the bottom–up processes emerge from the biological levels and give birth to the top–down processes.

the cortico-centric view of the psyche which considers the cortex as the center of consciousness. Solms rather suggests that there is a primary and affective[13] form of consciousness closely connected with the reticular system[14] which exists prior to the cerebral cortex[15]. Thus, Solms argues that consciousness depends initially on logics relating to the Freudian id rather than the ego[16]. As for the cortex, its essential function is not to produce consciousness, but to "stabilize" objects of perception, and it is "merely a repository of memory images" (Solms, 2018, p. 6). Mental representations may thus attain preconscious and conscious processes when they are transformed by the cortex into a material sufficiently stable to become the object of working memory. To put it differently, for Solms, "The essential function of the cortex" is to generate "stable, representational 'mental solids' that, when activated (or 'cathected') by affective consciousness, enable the id to picture itself in the world and to think" (2013, p. 14). The cortex would thus contribute to the emergence of a "space of representational memory" from which free association could be deployed.

Solms also supposes that "free energy minimization is the basic function of homeostasis" and that "the functions of homeostasis and consciousness are realized physiologically in the very same part of the brain" (Solms, 2018, p. 10). Consciousness would then be "an extended form of homeostasis" conducting to a specific functional organization which would represent an adaptive advantage. In Solms (2013)' model, primary processes appear

(Solms and Panksepp, 2012, p. 151). The origins of consciousness thus appear at an anatomical subcortical level which corresponds to seven primary processes shared by all mammals: seeking, lust, care, play, rage, fear, and panic. Therefore, Solms and Panksepp argue that "all consciousness ultimately derives from upper brainstem sources" (p. 163) and that consciousness is, metapsychologically, "generated in the id" (p. 164). Consequently, consciousness initially emerges distinctly from reflexive processes and the anoetic level probably emerged before the noetic and autonoetic levels.

[13]Similar claims have also been brought forward by Damasio (2010) who proposes a distinction between the proto-self, the core consciousness, and the extended consciousness. See Rabeyron (2016) for more details concerning parallels between Damasio's model and psychoanalytical models of consciousness.

[14]That is, the neurological structure of the brainstem, influenced by somatic and emotional stimuli responsible for muscle tone and vigilance states.

[15]Solms relies on data originating from neuroscientific research on hydranencephaly, a condition marked by the destruction of the cerebral cortex *in utero* (2013, p. 10), where one can nonetheless notice all signs of a primary form of consciousness.

[16]Solms and Panksepp (2012) propose that secondary processes derive anatomically from the stabilization of representations by the cortex. The tertiary processes Panksepp describe correspond to Freud's secondary processes for they use "word-representations" whose "symbolic nature enables to represent abstract relations between the concrete objects of thoughts" (p. 166).

to belong to a first form of consciousness, characterized mainly by affects, and preceding a secondary form of consciousness whose function is to stabilize mental objects. In other words, the transition from primary processes to secondary processes would correspond to the way in which free energy becomes bound by secondary processes, thus permitting the stabilization of mental representations and their access to a secondary or reflexive form of consciousness (Solms, 2013). But how might this transition from the primary affective consciousness to the secondary consciousness, from free energy to bound energy, arise? And what is the influence of this transition on our understandings of free association?

According to Carhart-Harris and Friston (2010), this transition emerges thanks to the "default mode network" (DMN). They suppose that the DMN is consistent with Freudian ideas of the ego that could take part into this transition from primary to secondary processes. The DMN defines a network that develops during childhood and connects several anatomical zones remaining active during the resting state – notably the medial temporal lobe, the medial prefrontal cortex, the posterior cingulate cortex, the precuneus, and other neighboring regions of the parietal cortex (Buckner et al., 2008)[17]. It consumes more energy than any other areas of the brain, a fact that signals a high associative density between these other areas. For Carhart-Harris and Friston, the activation of the DMN also corresponds to a decrease in the activity of the lower levels of organization, which suggests that it serves to modulate internal and external inputs or to suppress prediction errors (the free energy stemming from lower levels of mental functioning). The DMN is mainly engaged in higher mental operations, such as meta-cognition and reflexivity, as shown by several imaging protocols (Carhart-Harris and Friston, 2010). Spontaneous oscillations in the posterior cingulate cortex, particularly in the alpha of 8–13 Hz, are a neurological marker of the DMN's functioning that Carhart-Harris and Friston further link to a possible work of integration by the ego (2010, p. 2). Lastly, the activation of the DMN is inversely proportional with the attention system[18] and its activity appears to decrease with age as well as in people with attention deficit disorders.

From these different elements, Carhart-Harris and Friston hypothesize that the functioning of the DMN offers a neurobiological equivalent to Freud's ego. More precisely, according to Friston (2009), conscious activity, linked to the processes of the DMN, would constitute a temporary measure of adaptation between the brain and the environment. The brain must attempt to "correct" any discrepancy between the internal model of reality and the external reality. Thus, the fundamental aim of the DMN would be to limit its activity by seeking an "automatic" mode that would minimize the necessary adjustments between internal reality and external reality. Cieri and Esposito (2019) also suppose that "the DMN seems to play the same function of mediation attributed by Freud to the ego and some authors have spoken about *Default Self* in order to define the DMN as a kind of biomarker of the Self" (p. 6). They add that "the DMN is consistent with ego functions and with its target of containing free energy levels of underlying structures, a function of the secondary process. The result is a top–down hierarchy of DMN which aims to reduce the free energy associated with the Freudian primary process" (p. 12). Dimkov (2019) suggests an alternative view in which DMN is co-activated with Centre-Executive Network during regression processes. From his point of view, "DMN appears to function as a third thought process, an intermediary process between the primary and the secondary ones" (p. 170).

Solms (2013) argues that this work of articulation and prediction operates during the transition from primary to secondary processes, or from affective to stabilized consciousness[19]. The world becomes more organized and "predictable" as the effects of surprise diminish. The main function of consciousness is then to carry out this work of prediction through the affects self-informing the subject regarding the relevance of its generative model. The primary relation to the world is thus an affective relation intrinsically linked to the pleasure principle. Secondary consciousness serves to "re-work" unrepresented affects arising from a painful discrepancy between internal and external worlds. When the effects of surprise disappear, this form of consciousness is no longer necessary, in the same way that the dancer does not need to reflect upon the movements practiced a thousand times. Linguistic systems (word-representations) allow the subject to regulate these primary affects, opening the way to forms of associativity obeying different principles. Signifiers will thus participate in secondary processes by adding an additional "delay" and reshaping associative processes depending on structural laws of language (Lacan, 1966). Thus occurs a transfer from the primary associative logics to the linguistic apparatus, from affects to word-representations (see **Figure 2**). The latter, as Roussillon et al. (2007) suggest, will nevertheless keep the "trace" of this passage through the body in the form of a particular "corporeity" or "materiality" of language, showing that one form of associativity does not make the previous one disappear[20]. Free association – expressed linguistically – thus keeps the influence of the subject's overall psychic functioning, which explains its essential function in the exploration of the patient's psyche. For the analyst, it is the equivalent of the biologist's microscope in that it allows for an "inside view" of psychic functioning as it occurs.

[17]It should be mentioned here that the variation in global cerebral activity from focus on a task to the resting state is only 5% (Raichle and Snyder, 2007), which may suggest that psychic work can largely operate independently of the environment.

[18]The attention system is not unified. It is divided between *salient system* and *dorsal attention system*. Their hierarchic relation is established at the higher level of the DMN, while they themselves occupy intermediary levels.

[19]The transition from an affective to a stabilized form of consciousness is described by Solms as the transition from primary to secondary processes (Solms, 2013) or from primary–secondary processes to tertiary processes (Solms and Panksepp, 2012) depending on which model he relies on (mainly Freudian or Panksepp models). As in Solms (2013), we will use the distinction between primary and secondary processes to describe this transition in the following pages.

[20]Regarding linguistic developments stemming from bodily schemas, see notably Lakoff and Johnson (2003). Language emerges from metaphoric and metonymic processes – as Lacan (1966) earlier suggested – derived from physical experiences. Each level of the generative model is thus influenced by the logics and peculiarities of the lower levels.

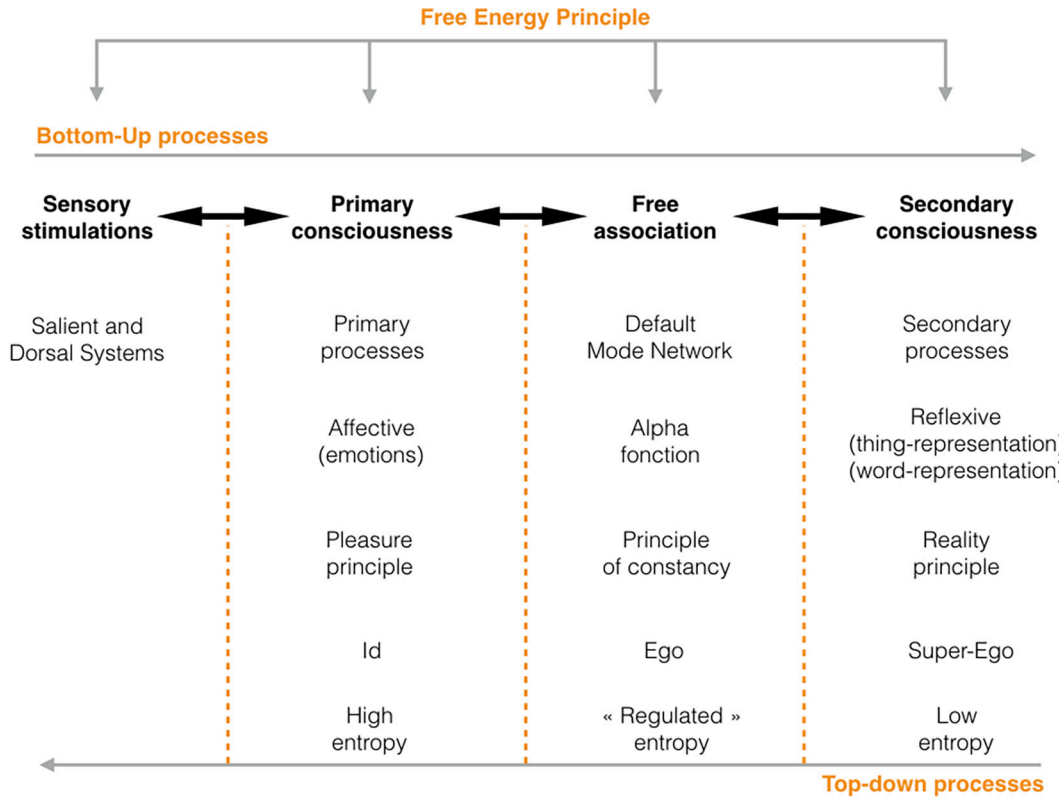

FIGURE 2 | The FEP applies at the different levels of mental functioning from sensory stimulation to secondary consciousness. We have here indicated the major neurobiological processes, the type of processes and principles, the equivalents in the Freudian topography, as well as the level of entropy at each level of functioning. Following Dimkov's hypothesis, we suggest that the DMN is a process of articulation between primary and secondary processes rather than the expression of the functioning of secondary consciousness. Free association thereby appears as the expression of the work of psychic integration carried out by the DMN. The following diagram does not pretend to represent a "scientific" model of psychic functioning but rather offers a general representation of the logics of functioning of the psyche and the intersections of neuroscience and psychoanalysis.

The brain thus appears to have evolved in order to simulate its environment and diminish the effects of surprise thanks to a Bayesian model. Free association can be considered as an echo of this process insofar as it reflects the functioning of psychic reality, itself constructed through a constant relation with the environment virtually simulated via a generative model (Hopkins, 2016). The intrapsychic associativity thus preserves the trace of the environment –an "externalized" associativity – to which the subject has been confronted.

FREE ASSOCIATION AND THE ENTROPIC BRAIN

Carhart-Harris extends this understanding of associativity with the "entropic brain" theory, which refers to the degree of organization or uncertainty of conscious states (Carhart-Harris et al., 2014). For Carhart-Harris, the hierarchical structures of the brain are situated within a continuum depending on different levels of organization. The primary consciousness described by Solms corresponds to a higher degree of entropy, for it is less "meticulous" in its relation to the world and is highly malleable. The secondary consciousness, on the other hand, works to diminish high entropy levels resulting from the primary consciousness by organizing and constraining cognition. Thus, primary consciousness is more "entropic" and flexible than secondary consciousness, which presents a higher degree of organization and a lower degree of entropy[21]. When the relation to the environment becomes a source of uncertainty or puzzlement, the subject has to "contain" this uncertainty[22]. The subject can then react in different ways: for example, "magical thinking" will interpret the world according to one's desires (the pleasure principle) when a high level of entropy "overflows"

[21] Hobson et al. (2014) and Hopkins (2016) also argue that primary processes form an "innate virtual reality" that characterizes dream processes and a high degree of entropy.

[22] The ability to contain uncertainty is an essential component of Bion (1965) theories. The latter argues that symbolization processes require a "negative capability" (a term borrowed from the poet Keats). Within psychoanalysis and psychodynamic therapies, this expression signifies that, to integrate an experience, the subject must first be able to deconstruct it in order to rebuild it differently. This process of deconstruction and construction requires the subject to tolerate uncertainty, represented by the negative capability (Rabeyron, 2018). In other words, the deployment of the association necessitates the capacity to accept uncertainty in order to catalyze symbolization.

the secondary processes. Depressive states, on the other hand, will demonstrate a difficulty in balancing the uncertainty arising from primary levels of psychic functioning[23]. In such states, neuroimaging has revealed a hyper-activation of the DMN, a consequence of hypertrophied introspection and a desperate attempt of the ego to control the entropy stemming from primary processes. Carhart-Harris describes this movement through the theory of "self-organized criticality," which shows how a complex system develops specific properties when disturbed to a critical extent by a new energy input (Carhart-Harris et al., 2014). In the narrow transition zone between the extreme positions of chaos and order, three properties will emerge: metastable or transiently-stable states, a sensitivity to perturbation, and a propensity for cascade-like processes called "avalanches." These could find their correlate in the functioning of the ego and psychopathological expression. For example, avalanche processes could lead to psychotic collapse showing how the ego is suddenly unable to internalize new energy input[24].

Carhart-Harris et al. (2014) have experimentally studied these variations of organization and associativity through psychedelics – particularly psilocybin – and revealed that the latter produces a prototypical primary state of consciousness with high entropy. Psilocybin alters consciousness through a disorganization of cerebral activity, which translates into a significant decrease in the activity of key brain areas connected to the DMN. Psychedelics can thus generate profound states of insight concerning the self, often referred to as an oceanic feeling (Freud, 1930) of dissolution of the ego and its borders. The phases of paradoxical sleep, initial and acute psychotic periods, and certain epileptic states also seem to engender regression to primary consciousness. Thus, as Freud suggested, dreams and psychoses probably pertain to primary forms of consciousness (also dominating in infancy), while meta-cognition would develop only secondarily (on this topic, see also, Hopkins, 2016).

In sum, a distinction emerges between two main states of cognition, the first being characteristic of the state of consciousness of the adult, and the second, present in infancy, reappearing through mechanisms of regression. These two states of consciousness are related to certain frequencies of neuronal activity, in particular the power of alpha waves correlated with reflexive activity (Carhart-Harris et al., 2014). Certain cerebral rhythms correspond to a decrease in entropy due to an increase in the exchange of information between neural networks. The use of psilocybin, in particular, induces a decrease in the activity of the alpha spectrum, thus resulting in a subjective feeling of disintegration. Under the effects of such psychedelics, the brain behaves more randomly, its hierarchical functioning becomes anarchic, and the associativity becomes more flexible, regressing to primary modes of functioning[25]. The therapeutic effects of psychedelics could therefore rest on an "extreme" form of symbolization[26] different from the usual, more "attenuated" symbolization found in psychodynamic psychotherapy and psychoanalysis. The deployment of free association and the passage through high entropic states would allow for a necessary relaxation of the psyche, thereby reviving processes of symbolization. Examining free association therefore seems crucial for understanding psychic integration between internal and external worlds.

FREE ASSOCIATION, SYMBOLIZATION, AND PSYCHOANALYTICAL PRACTICES

To what extent do these theoretical models of free association resonate with, and affect, clinical practice? While the works discussed above essentially focus on normal cognitive functioning, clinicians generally work with the failures of the "associative machine" or try to catalyze symbolization[27] processes thanks to free association[28]. When the clinician asks the patient to verbalize "everything that comes to mind," he or she intends to help the latter "untie" the free association that leads to a subjective and complex mix of sensations, emotions, images, words and memories. The patient will then use spontaneously the different "languages" at his or her disposal,

[23]Depression appears more precisely as a mechanism of withdrawal and disinvestment from the environment that allows the subject to reduce the states of excitement stemming from interactions with the outside world. The depressed subject is frequently confronted with psychic conflicts increasing free energy. The withdrawal of the depressed subject thus offers a solution to reduce the entropy. Yet, this solution has the disadvantage of preventing the implementation of adaptive strategies. A withdrawal process of a different nature also emerges in psychosis (De Masi et al., 2015).

[24]In psychosis, we witness the disintegration of the psychic "membrane" separating conscious and unconscious processes when the "disobjectalizing function" of the death drive described by Green pervades the psyche (Mellor, 2018). The formation of the symbol, an essential component of psychic functioning, is no longer operational and the words are treated as things within the "symbolic equation" described by Anna Segal (1957).

[25]While it is important, from a theoretical perspective, to distinguish these two modes of functioning and their neurobiological correlates, one must also bear in mind that they probably work in parallel. Kahneman (2011) research indeed demonstrates that System 1 and System 2 often work in concert.

[26]As their name indicates, extreme forms of symbolization designate the extreme and unusual expressions of the subject's psychic integration and transformation. They emerge, notably, during altered states of consciousness (Cardeña and Winkelman, 2011) and various anomalous or exceptional experiences (Rabeyron et al., 2010; Rabeyron and Loose, 2015). In relation to Friston's theories, extreme forms of symbolization may correlate with a state of global reorganization of the generative model.

[27]The notion of symbolization deserves some precisions (see also, Roussillon, 2015). The word "symbol" originates from the Greek *sumbolon* and the verb *sumballesthai*, which signifies "bringing together." In ancient Greek, the symbol defines a broken shard of pottery whose parts are given to a contractor, the latter ensuring the authenticity of the "contract" by reassembling the pieces. More specifically, the term symbolization designates the operation through which one thing represents another. In a clinical sense, the symbolization defines the various steps that permit the subject to transform "primary psychic matter" into a state of reflectivity and subjectivity. In other words, the symbolization designates the many ways the subject transforms its experience in order to integrate and appropriate it subjectively. From the perspective of psychopathology, a variety of sufferings may be understood as "blockages" of symbolization processes. This is why it is important to understand the modalities and logics of symbolization. We distinguish more precisely between primary or archaic symbolization, which Freud designates as thing-representation, and secondary symbolization, which Freud calls word-representation.

[28]Within therapy, these symbolization processes were developed through clinical encounters with borderline and narcissist personalities (Green, 1990). Such encounters require therapists and patients to work at very primary levels of psychic functioning which are hardly reached through verbal or linguistic expression. This is why the expression of associativity has been developed by Roussillon (2015) to underscore the importance of mediations for the patient's elaboration of traumas dividing the personality.

such as breathing, motion of the body, facial expressions, words and narratives to share his or her psychic life in the clinical setting[29]. The work of free association therefore follows the different modes of symbolization in order to share, integrate and transform the internal experience within the present therapeutic intersubjective relationship.

A primary form of free association concerns mainly the emotions emerging within the dialogue of therapy. This form of shared consciousness, or this "affective co-consciousness," relies on primary processes and arises from the clinician's regressive skills at the most primary levels. As Parat (2013) explains, "This forms the possibility of an inter-human relationship, which is established directly and regressively in a preverbal, ante-verbal mode, and where the affect of one echoes the affect of the other. [This relationship] is perhaps the only way to allow for the approach and mobilization of the elements, the sediments of the primary repression" (p. 171). For Parat, the clinician's position opens the way to a "basic transfer," an expression that approaches Stern (2004) "intersubjective sharing" or Christian David (1992) "accompanying activity." This sharing of affects presents a first form of intersubjective and undifferentiated associativity that permits the release of a non-symbolized psychic "residue." As de M'Uzan (1994) work on "paradoxical thought" (*chimère*) and Widlöcher (1996) concept of "systems of co-thinking" demonstrate, this primary associativity can be particularly "permeable," as it is characterized by psychic transmissions from unconscious to unconscious (Evrard and Rabeyron, 2012). The psychotherapeutic dyad thus produces an "analytic third" (Ogden, 1994) which combines the indissociable thoughts of patient and clinician[30].

At a more elaborate level of psychic functioning, the passage through words, to form new signifying chains, breaks this primary and shared form of regression. In other words, there emerges a "secondary associativity" that requires the patient to come out of his or her state of regression and to integrate the experience at higher levels of functioning. This work operates more specifically through conscious activity and word-representation, and it reduces the entropy coming from lower levels of mental functioning. The patient may then deploy more elaborated and secondary levels involving the stabilization of mental objects, as Solms (2013) suggests. This work of stabilization may participate in the process of psychic integration as evidenced, for example, by the patient who suddenly becomes able to understand previously unintelligible parts of his or her experience (affects, behavior, etc.). It produces what Freud (1937) called a "construction" – or what Bion (1965) named a "selected fact" – that re-organizes experiences through the medium of speech. Here, perhaps in ways similar to the processes of "reconsolidation" of memory traces described by Alberini et al. (2013), the raw experience can then be treated again through the different levels of associative processes as the linguistic apparatus comes to regulate primary processes. Words then come to the rescue of the body and the unrepresented affect.

Free association thus emerges as an essential component of this work of symbolization at the intersection of primary and secondary processes. It permits the subject to diminish its investment in the external environment in order to increase attention to intrapsychic reality. Akin to the caterpillar metamorphosing in its cocoon, the patient can here safely elaborate the experiences that have not been integrated in the psyche (Rabeyron, 2019). Free association thereby augments the free energy that was, until then, contained by defense mechanisms such as repression and splitting. Free association increases the prevalence of primary processes, thereby occasioning regressive and hallucinatory states[31]. This regression to primary processes truly allows for psychic integration to occur when coupled with secondary processes as a complementary system necessary for the reflexive metabolization of the subjective experience. This requires a very particular psychic activity which underlies the effects of free association and which corresponds to what Bion (1965) calls the "alpha function" permitting the transformation of sensations and emotions into thinkable contents. Freud (1900) had already intuited this function in suggesting that the dream was necessary for the passage from primary to secondary processes. Bion (1965), however, demonstrated that the dream work is always present in the psyche. We "dream" both day and night insofar as we constantly need to transform our experiences into subjective psychic matter. The distinction between thinkable and unthinkable thoughts emerges in the passage through the alpha function which distinguishes, by a membrane, the conscious and unconscious processes.

The DMN could be a neurobiological equivalent of the alpha function described by Bion[32], a function that bridges various

[29]The verbal expression of free association emerges from other forms of associativity. Thus, the spontaneous free play of the baby (Pikler, 1962), which passes from one object to another in the exploration of the surrounding world, appears as a preform of free associativity. We first learn to play – to associate – with objects before we begin to play with words. Yet, the verbal mode of free association does not entail the disappearance of this first mode of free association. This explains why a non-verbal form of free association emerges in the clinical setting and why the therapeutic relationship must be attuned to this other form of associativity.

[30]This process finds echo in the work of Friston and Frith (2015) examining how the generative model can be improved, thanks to synchronization processes, as seen at a behavioral level by the simulation of birdsongs. These synchronization processes are particularly complex because "we are trying to infer how our sensations are caused by others, while they are trying to infer our behaviors" (p. 1). In this regard, "communication facilitates long-term changes in generative models that are trying to predict each other" (p. 12). If these two generative models are close enough in their functioning and share a collective narrative (in psychotherapy, the same language and the same culture), they will progressively become attuned and each of them will develop a generative model with better predictability (and consequently a reduction of free energy). During this synchronization process, hidden states will emerge that belong to *both* birds which could correspond to what is described notably by Ogden (1994) or Green (2002) as the analytical third and tertiary processes. In other words, this research suggests that synchronization processes are essential in order to improve generative models, a discovery that may underline a fundamental aspect of the psychotherapeutic process.

[31]César and Sara Botella have approached the question of psychic figurability through regression and hallucinatory states as a "work of transformation directed toward the implementation of a psychical intelligibility and heterogeneity" (Botella, 2006, p. 5). See also Botella and Botella (2001).

[32]It is not, however, a question of reducing the alpha function to a neurobiological network, but of supposing that properties of psychic functioning arise from this level of biological organization, while being irreducible to biological processes. On this epistemological topic, see the analyses of Connolly and van Deventer (2017).

levels of psychic integration and more particularly the primary and secondary processes[33]. The work of psychic integration requires both the dream regression and the elaboration of past experiences. Bion (1965) model also insists on the idea that the alpha function results from the integration of the alpha function of the mother. This function involves three factors: daydreaming, diffraction-synthesis and contained-container. The first of these factors – daydreaming – occurs as the mother takes care of the child and is the prototype of the alpha function. The work of articulation between psyche and soma, between the unconscious and consciousness, therefore emerges from the early intersubjective relation between the baby and the mother. This long and complex process may explain why subjectivity, from a psychoanalytical point of view, takes many years to emerge in the human being.

The emergence of subjective experience through an intersubjective process has been recently examined by Holmes and Nolte (2019) from the perspective of the Bayesian brain. They note that the development of the generative model occurs through the "borrowing" of the maternal brain and suggest that "this borrowed brain model introduces a vital interpersonal dimension to the Bayesian process" (p. 4). The baby thus internalizes the experience of maternal care in order to reduce the entropy: "these embodied gestures present a model of the infant from the caregiver's perspective helping the child to integrate primary sensory signals [. . .] into regularities of emotional and interpersonal consequences" (p. 4). Within psychodynamic therapies, a similar process emerges as the subject develops its capacity for psychic integration through the intersubjective relationship with the therapist. This relationship fosters the resurgence of a "we mode" (Frith, 2012) in which "two heads are better than one" given that the other can "know our self better than we can know ourselves" (p. 5)[34]. Thus, "one of the roles of psychotherapy is to reactivate this process" (p. 5) through the deployment of free association as the expression and elaboration of the intrapsychic dynamic. Free association might thereby emerge as the joint connection of two Bayesian brains progressively leading, through their synchrony, to the dissolution of boundaries. In this manner, the "therapeutic duet for one helps bind potentially disruptive free energy in creative ways" (p. 6).

Early traumatic experiences like, for instance, what Winnicott (1963) calls "primitive agonies" are not integrated because they induced too high levels of entropy and thereby could not be "bound" by the psyche for they were not sufficient sources of pleasure[35]. This failure of integration might lead to mechanisms beyond the pleasure principle, such as the repetition compulsion. The analytic work creates a regression to primary processes within the safe environment of therapy, which permits to deconstruct the cleavage resulting from such early traumatic experiences. Through the practice of free association, the patient may affectively experience these previously unmetabolized agonies. The regression to primary levels would also emerge through daydreaming[36] simultaneous with the free association. Free association would thus connect primary and secondary processes through the modalities of psychic integration permitting the renewal of symbolization processes.

It is perhaps in the crossing from primary to secondary processes that therapeutic gains are the most significant. Following the large body of research already developed concerning primary intersubjectivity and transmodal processes (Stern, 2000; Trevarthen and Aitken, 2001; Beebe et al., 2016), we could call this process "symbolizing transmodality." This notion relates to the way in which the psychotherapy allows for the associative transfer between the various forms of symbolization. It results from an intersubjective associativity, as it emerges from the relationship developed between clinician and patient. It explores the primary and preverbal modes of communication involved in the mother-baby relationship, including sequences of motions of the body, rhythms of speech, tone, voice, sounds, facial expressions, etc. (Stern, 2000). The symbolizing transmodality transforms what the subject tries to explore through another sensory aspect. This passage permits the subject to "restore" the symbolization process by using a different sensorial modality. Its function is to metaphorize the inner experience as it moves from the most primary and unconscious forms to the more secondary and conscious processes.

From this point of view the analytic session forms a containing space for an increase in free energy allowing the subject to safely make prediction errors and confront surprise effects. Hence this astonishing paradox, as already noted by Reik (1936), of the necessity for patients, as well as clinicians, to preserve the ability to be surprised during therapy[37]. Through their echoing – and their own negative capability (Bion, 1965) – clinicians will favor the effects of surprise in patients. In Friston's model, the effects of surprise are usually avoided by the psyche because they signify a gap between the internal and external worlds. Within psychoanalytic therapies, however, psychic mechanisms of "surprise" are required. For example, transference can be considered as a prediction error since the subject "confuses" the clinician with the parental imago. This confusion nonetheless

[33] Green (1995) notably calls "tertiary processes" the back and forth movements, or the binding processes, between primary and secondary processes, thus complicating Freud's binary model by emphasizing its interconnections (Green, 2005). In this perspective, both the alpha function and the DMN appear to fall under tertiary processes.

[34] In the same way that we cannot see our own face directly – while the other can – our own psychic reality seems, in certain respects, more accessible to others. For instance, others are generally better placed to tell us if we have some unusual element on our face. It is probably similar with psychotherapy, which might explain why, from an evolutionary perspective, the intersubjective relation is more efficient for the auto-representation of processes.

[35] This hypothesis comes from Freud (1920) who supposes that an experience needs enough pleasure to be integrated in psychological reality. A traumatic experience (leading to PTSD for example), if it is too painful, will not be integrated. From the FEP perspective, we can consider that the traumatic experience induces too much entropy – or excitation – to be integrated by the generative model of the brain. It will then be cleaved from the functioning of the generative model.

[36] Much neuroscientific research has examined Random Episodic Silent Thinking (REST) (Andreasen et al., 1995) and Mind-Wandering (Mooneyham and Schooler, 2013). While we may not address these here, they also present interesting parallels with states of daydreaming as understood in psychoanalysis.

[37] In psychoanalysis, the dynamic of transference is therefore understood to generate effects of surprise that can later lead to positive outcomes. This is, notably, an important aspect distinguishing psychoanalysts from cognitive and behavioral therapists, particularly within the medical field, where the clinician will work – at the opposite – to avoid or diminish the effects of surprise.

gradually allows the subject to refine its own internal model by managing to differentiate the clinician from this projection[38]. Throughout the sessions, effects of surprise may thus emerge as experiences of pleasure[39], for they can take the shape of sudden awareness or "eureka moments" leading to a improvement of the generative model. Such experiences reveal a form of free association and creativity[40]. They allow the patient to organize a set of internal representations through an externalized object supporting the projection of internal associativity. An encounter with an external object or an Other – whose properties favor processes of symbolization – produce an original subjective experience. The initial experience is thus transferred into the object and allows the subject to benefit from the symbolizing transmodality process[41].

CONCLUSION

During psychoanalytic and psychodynamic therapies, the patient passes from one idea to another and deploys a signifying chain composed of affects and representations using both verbal and non-verbal forms of expression. This free association process is an essential component of psychoanalytic practices and relies on complementary functions as illustrated in **Figure 3**. First, free association lets the subject express its intrapsychic world through increased focus on the internal experience and decreased focus on the environment. It allows for the exploration of intrapsychic reality – as a virtual reality generator (Hopkins, 2016) – by both the patient and the therapist, for the latter is also in a specific state of mental free association. Akin to dreams, free association also allows for the emergence of latent contents related to significant and mysterious elements of the subject's psychic life. In this sense, it permits one to recognize the traces of traumatic experiences having induced "permanent disturbances of the manner in which the energy operates" (Holmes and Nolte, 2019, p. 6) as well as the traces left by the repression of certain drives[42]. Thus, as Bion (1962, 1965) suggests, the dream-like state that accompanies the free association helps the patient to transform non-thinkable thoughts into thinkable thoughts. This work of psychic integration through free association relies on a dynamic of "un-translation-translation" (Laplanche, 1987) in order to foster a more global and coherent elaboration of psychic experiences. Thus, free association becomes an essential tool for the synthesis of the ego. It refines the subject's reflexivity as the subjectivation process develops and fosters symbolization of unelaborated traumatic experiences.

FIGURE 3 | The different functions of free association.

As proposed in this paper, this psychoanalytic understanding of free association also shares a number of theoretical parallels with contemporary neuroscientific models. As Scarfone (2018) suggests, "the contemporary usage of Helmholtzian ideas in brain science does indirectly support and justify recurring to free association in psychoanalysis" (p. 468). The activity of thought appears to correspond to a biological and psychological organization at primary and secondary levels whose cognitive correlates may be found in Kahneman (2011) System 1 and System 2. For Freud, a two-level division of mental processing (primary and secondary processes) is associated with the passage of "free energy" to "bound energy" according to a complex and hierarchical organization. In Solms' model of the psyche, there is indeed first a primary form of consciousness which functions mainly through affects and an unbridled associativity. The second psyche's main function is to form a coherent representation

[38]To suggest a simple metaphor, just as, in skiing, one may only progress by learning to fall, in psychoanalysis, one must be allowed to make "mistakes" otherwise unacceptable in daily life.

[39]Perhaps the surprise might also induce pleasure when an increase in entropy leads to a reorganization of the generative model. On the other hand, when the degree of surprise is too great, we might assume that it cannot induce a reorganization and the experience therefore becomes unpleasant. Thus, as Holmes and Nolte (2019) suggest, "all depends on decoupling – introducing a degree of play into the Bottom Up/Top Down surprise –minimizing articulations of everyday life" (p. 8). The tact shown by the clinician might be related to this ability to induce an appropriate amount of surprise and to avoid the use of defense mechanisms as means of curbing an excessive entropy.

[40]For example, in the field of science, Newton's discovery of the laws of gravity under an apple tree, or Einstein's theorization of special relativity after repeated observations of clocks and trains at Bern station, offer examples of this creative associativity that comes from transmodal experiences.

[41]Clinical work with children, where the symbolic transmodality expresses itself essentially through play, offers an illustration of this process. Indeed, the work of "stabilization" of mental objects may be sub-operative, leading the child to "stick" the self to objects, as Winnicott (1958) suggested through the notions of "transitional space" and the "found-created." Hence the complexity of the clinic with children, where the very first session will present the latter's preoccupations, but in a condensed or compressed way. The child's first play, in particular, condenses the associations relating to the causes of the child's suffering, which often become intelligible only through multiple sessions.

[42]From this point of view, what is most essential is not what the patient says, but what the patient does not say because the negation emerges as a consequence of defense mechanisms and repression.

of the world using secondary and tertiary processes (Solms and Panksepp, 2012). Each of these levels works to limit the effects of surprise and disorganization. The transition between these levels of consciousness and their modes of associativity could emerge through the DMN whose purpose is to integrate and organize internal and external information. As Cieri and Esposito (2019) suggest: "Freudian constructs of the primary and secondary processes seem to have neurobiological substrates, consistent with self-organized activity in hierarchical cortical systems, and Freudian descriptions of the ego are consistent with the functions described of the DMN with its reciprocal exchanges with subordinate limbic and paralimbic brain systems" (p. 12). The role of psychoanalytic therapy, therefore, appears to be the reestablishment of this integrative work.

To conclude, we propose two further avenues of research that may be useful in orienting future work about free association. First, why is there a need for "associative transference" (from self to object or from self to other) for the subject to metabolize and integrate certain psychic contents? Current research about synesthesia – from the Greek "sunaisthesis" or simultaneous perception – may answer this question by addressing the unusual association of various senses. For instance, according to such research, a subject – notably individuals with autism (Neufeld et al., 2013) – may visually perceive colors associated with specific musical notes, numbers, or alphabetical letters. An immediate and automatic association from one sensory perception to another may thus emerge, which cerebral development would normally inhibit (Ward, 2013). Such form of "primordial associativity" may remain partly present in the psyche, while the symbolizing transmodality would emerge as its vestige. Innovative research joining, for instance, clinical and cognitive paradigms related to phenomena of synesthesia may thus lead to better understandings of free association and symbolization processes.

A second avenue might concern the limits of epistemological approaches to psychic states as they emerge thanks to free association. Psychoanalysts suggest that free association does not solely operate as a work of synthesis of the ego, but also as a work of disintegration of subjective experience. Thus, Scarfone (2018) reminds that "the lysis part of analysis literally means unbinding" (p. 473). The dissolution, or unbinding, work of analysis might be compared metaphorically to the nuclear fission defining the release of energy produced by the division of a heavy nucleus. Similarly, the work of analysis might release unbound energy through the free association process. For Barratt (2017), "herein lies the distinction between our discipline and all the therapies that prioritize the discourse of synthesis and integration" (p. 48). Furthermore, one may never represent the entirety of unconscious psychic activity, as free association is also a contact of the unknown. Thus, from an ontic perspective, "the praxis of free association effects an *ontic* change – a transmutation in the being of the subjective – which is not going to be explainable epistemologically." (p. 49). For the same reason, Scarfone (2018) claims that "the repressed unconscious will never be fully transmuted into ego" (p. 476) and a part of the subjective experience will always remain as "non-representation" (David, 1992). Thus, there would be a risk in trying to explain or "rationalize" everything that is proposed by the patient[43] and it could lead to pseudo-advances in the therapeutic process (Stern, 2011). Barratt (2017) also underlines that, from a clinical point of view, "the implication is that whereas one might become able to listen to the voicing of the repressed, through the praxis of free association, it is not to be assumed that the meaningfulness of the repressed is entirely translatable into the languages of representation" (p. 42). Something might have to remain obscure – what Freud calls the dream's navel – and untranslatable into representable thoughts. Thus, Barratt (2017) continues, one part of the "repressed unconscious necessarily remains unknown (that is, unrepresentable), precisely because it involves impulses that are ontologically different from the meaningfulness of representationally" (p. 42). Such reflections might suggest the necessity to examine how far the connections between neuroscience and psychoanalysis might go and to what extent there may also be, for epistemological reasons, fundamental differences in regards to the knowledge emerging from these two complementary domains.

AUTHOR CONTRIBUTIONS

TR wrote the first draft of the manuscript. CM helped to translate and improve the quality of the writing of the manuscript.

[43] Bion (1965) calls −K the link which does not permit the authentic integration of an experience. From a Bionian perspective, free association is a means to establish a contact with O a an unattainable Real that underlies biological and psychic dimensions.

REFERENCES

Alberini, C. M., Ansermet, F., and Magistretti, P. (2013). "Memory reconsolidation, trace reassociation and the Freudian unconscious," in *Memory Reconsolidation*, ed. C. M. Alberini (New York, NY: Academic Press), 293–310.

Andreasen, N. C., O'Leary, D. S., Cizadlo, T., Arndt, S., Rezai, K., Watkins, G. L., et al. (1995). Remembering the past: Two facets of episodic memory explored with positron emission tomography. *Am. J. Psychiatr.* 152, 1576–1585.

Anzieu, D. (1974). Le moi-peau. *Nouv Rev. Psychanal.* 9, 195–208.

Aulagnier, P. (1975). *La violence de L'interprétation (2003 Edn)*. Paris: Puf.

Barratt, B. B. (2016). *Radical Psychoanalysis: An Essay on Free-Associative Praxis*. London: Routledge.

Barratt, B. B. (2017). Opening to the otherwise: the discipline of listening and the necessity of free−association for psychoanalytic praxis. *Int. J. Psychoanal.* 98, 39–53. doi: 10.1111/1745-8315.12563

Beebe, B., Messinger, D., Margolis, A., Buck, K. A., and Chen, H. (2016). A systems view of mother-infant face-to-face communication. *Dev. Psychol.* 52, 556–571. doi: 10.1037/a0040085

Bion, W. (1965). *Transformations: Passage de l'apprentissage à la croissance* (2002 Edn). Paris: PUF.

Bion, W. R. (1962). *Aux Sources de L'expérience* (2003 Edn). Paris: PUF.

Bollas, C. (2008). *The Evocative Object World*. London: Routledge.

Botella, C. (2006). *Rêverie-Rêverie et Travail de Figurabilité. Table Ronde, Débats Sans Frontières*. Paris: Société Psychanalytique de Paris.

Botella, C., and Botella, S. (1990). "La problématique de la régression formelle de la pensée et de l'hallucinatoire," in *La psychanalyse: Questions Pour Demain*, ed. I. Schimmel (Paris: PUF).

Botella, C., and Botella, S. (2001). *La Figurabilité Psychique (2007 Edn)*. Paris: in press.

Brun, A. (2014). Médiation thérapeutique picturale et associativité formelle dans les dispositifs pour enfants avec troubles envahissants du développement. *La Psychiatr. Enfant* 57, 437–464.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011

Cardeña, E., and Winkelman, M. (2011). *Altering Consciousness: Multidisciplinary Perspectives*. London: Praeger.

Carhart-Harris, R. L., and Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of freudian ideas. *Brain* 133, 1265–1283. doi: 10.1093/brain/awq010

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., and Tagliazucchi, E. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Front. Hum. Neurosci.* 8:20. doi: 10.3389/fnhum.2014.00020

Cieri, F., and Esposito, R. (2019). Psychoanalysis and neuroscience: the bridge between mind and brain. *Front. Psychol.* 10:1790. doi: 10.3389/fpsyg.2019.01983

Connolly, P., and van Deventer, V. (2017). Hierarchical recursive organization and the free energy principle: from biological self-organization to the psychoanalytic mind. *Front. Psychol.* 8:1695. doi: 10.3389/fpsyg.2017.01695

Damasio, A. R. (2010). *L'autre Moi-Même les Nouvelles Cartes du Cerveau, de la Conscience et Des Émotions*. Paris: O. Jacob.

David, C. (1992). *La Bisexualité Psychique*. Paris: Payot.

De Masi, F., Davalli, C., Giustino, G., and Pergami, A. (2015). Hallucinations in the psychotic state: psychoanalysis and the neurosciences compared. *Int. J. Psychoanal.* 96, 293–318. doi: 10.1111/1745-8315.12239

de M'Uzan, M. (1994). *La Bouche de L'inconscient*. Paris: Gallimard.

Dienes, Z. (2011). Bayesian versus orthodox statistics: which side are you on? *Perspect. Psychol. Sci.* 6, 274–290. doi: 10.1177/1745691611406920

Dimkov, P. R. (2019). Large-scale brain networks and freudian ego. *Psychol. Thought* 12, 14–27. doi: 10.5964/psyct.v12i2.328

Donnet, J. L. (2012). Le procédé et la règle: l'association libre analytique. *Rev. Fr. Psychanal.* 76, 695–723.

Eschel, O. (2006). Where are you, my beloved? On absence, loss, and the enigma of telepathic dreams. *Int. J. Psychoanal.* 87, 1603–1627. doi: 10.1516/7gm3-mldr-1w8k-lvlj

Evrard, R., and Rabeyron, T. (2012). Les psychanalystes et le transfert de pensée: enjeux historiques et actuels. *Evol. Psychiatr.* 77, 589–598. doi: 10.1016/j.evopsy.2012.05.002

Freud, S. (1895). *Project for a Scientific Psychology. Standard Edition of Complete Works*, Vol. I. London: Hogarth Press.

Freud, S. (1900). *The Interpretation of Dreams. Standard Edition of Complete Works*, Vol. IV-V. London: Hogarth Press.

Freud, S. (1901). *The psychopathology of everyday life. Standard Edition of Complete Works*, Vol. IV-V. London: Hogarth Press.

Freud, S. (1905). *Fragment of an Analysis of a Case of Hysteria. Standard Edition of Complete Works*, Vol. VII. London: Hogarth Press.

Freud, S. (1909). *The Rat Man. Standard Edition of Complete Works*, Vol. X. London: Hogarth Press.

Freud, S. (1913). *On the beginning of treatment. Standard Edition of Complete Works*, Vol. XII. London: Hogarth Press.

Freud, S. (1915). *The Unconscious. Standard Edition of Complete Works*, Vol. XIV. London: Hogarth Press.

Freud, S. (1920). *Beyond the Pleasure Principle. Standard Edition of Complete Works*, Vol. XVIII. London: Hogarth Press.

Freud, S. (1930). *Civilization and its Discontents. Standard Edition of Complete Works*, Vol. XXI. London: Hogarth Press.

Freud, S. (1937). *Constructions in Analysis. Standard Edition of Complete Works*, Vol. XXIII. London: Hogarth Press.

Freud, S. (1941). *Findings, Problems, Ideas. Standard Edition of Complete Works*, Vol. XXIII. London: Hogarth Press.

Freud, S., and Breuer, J. (1895). *Studies in Hysteria. Standard Edition of Complete Works*, Vol. II. London: Hogarth Press.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface* 10, 20130475. doi: 10.1098/rsif.2013.0475

Friston, K., and Frith, C. (2015). A duet for one. *Conscious. Cogn.* 36, 390–405. doi: 10.1016/j.concog.2014.12.003

Frith, C. D. (2012). "Implicit metacognition and the we-mode," in *Paper perended at Workshop on "Pre-reflective and Reflective Processing in Social Interaction* (Cambridge: Clare College, University of Cambridge).

Georgieff, N. (2010). Psychanalyse, neurosciences et subjectivités. *Neuropsychiatr. Enfance Adolesc.* 58, 343–350.

Green, A. (1990). *La Folie Privée*. Paris: Gallimard.

Green, A. (1995). Note sur les processus tertiaires. *Rev. Fr. Psychanal.* 36, 151–155.

Green, A. (2002). *Idées Directrices Pour une Psychanalyse Contemporaine*. Paris: PUF.

Green, A. (2005). *Key Ideas for a Contemporary Psychoanalysis*. New York, NY: Routledge.

Hobson, J., Allan, Hong Charles, C.-H., and Friston Karl, J. (2014). Virtual reality and consciousness inference in dreaming. *Front. Psychol.* 5:1133. doi: 10.3389/fpsyg.2014.01133

Holmes, J., and Nolte, T. (2019). Surprise" and the bayesian brain: implications for psychotherapy theory and practice. *Front. Psychol.* 10:592. doi: 10.3389/fpsyg.2019.00592

Holt, R. R. (1962). A critical examination of Freud's concept of bound vs. *Free cathexis*. *J. Am. Psychoanal. Assoc.* 10, 475–525. doi: 10.1177/000306516201000302

Hopkins, J. (2016). Free energy and virtual reality in neuroscience and psychoanalysis: A complexity theory of dreaming and mental disorder. *Front. Psychol.* 7:922. doi: 10.3389/fpsyg.2016.00922

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* 58, 697–720. doi: 10.1037/0003-066x.58.9.697

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straud and Giroux.

Lacan, J. (1966). *Écrits*. Paris: Le seuil.

Lakoff, G., and Johnson, M. (2003). *Metaphors we Live by*. Chicago: University of Chicago Press.

Laplanche, J. (1987). *Nouveaux Fondements pour la Psychanalyse*. Paris: Puf.

Leuzinger-Bohleber, M., Kaufhold, J., Kallenbach, L., Negele, A., Ernst, M., Keller, W., et al. (2019). How to measure sustained psychic transformations in long-term treatments of chronically depressed patients: symptomatic and structural changes in the LAC Depression Study of the outcome of cognitive-behavioural and psychoanalytic long-term treatments. *Int. J. Psychoanal.* 100, 99–127. doi: 10.1080/00207578.2018.1533377

Lothane, H. Z. (2018). Free association as the foundation of the psychoanalytic method and psychoanalysis as a historical science. *Psychoanal. Inquiry* 38, 416–434. doi: 10.1080/07351690.2018.1480225

Magistretti, P., and Ansermet, F. (2010). *Neurosciences et Psychanalyse*. Paris: Odile Jacob.

Mancia, M. (2006). Implicit memory and early unrepressed unconscious: their role in the therapeutic process (How the neurosciences can contribute to psychoanalysis). *Int. J. Psychoanal.* 87, 83–103. doi: 10.1516/d43p-8upn-x576-a8v0

Méheust, B. (1999). *Somnambulisme et Mediumnité*. Paris: Les Empêcheurs de penser en rond.

Mellor, M. J. (2018). Making worlds in a waking dream: where bion intersects friston on the shaping and breaking of psychic reality. *Front. Psychol.* 9:1674. doi: 10.3389/fpsyg.2018.01674

Mooneyham, B. W., and Schooler, J. W. (2013). The costs and benefits of mind-wandering: a review. *Can. J. Exp. Psychol.* 67, 11. doi: 10.1037/a0031569

Neufeld, J., Roy, M., Zapf, A., Sinke, C., Emrich, H. M., Prox-Vagedes, V., et al. (2013). Is synesthesia more common in patients with Asperger syndrome? *Front. Hum. Neurosci.* 7:847. doi: 10.3389/fnhum.2013.00847

Ogden, T. H. (1994). The analytical third: working with intersubjective clinical facts. *Int. J. Psycho. Anal.* 75, 3–20.

Panksepp, J. (1998). *Affective neuroscience: The Foundations of Human and Animal Emotions*. Oxford: Oxford UP.

Panksepp, J. (2010). Affective neuroscience of the emotional BrainMind: evolutionary perspectives and implications for understanding depression. *Dialogues Clin. Neurosci.* 12, 533–545.

Panksepp, J., and Solms, M. (2011). What is neuropsychoanalysis? Clinically relevant studies of the minded brain. *Trends Cogn. Sci.* 16, 6–8. doi: 10.1016/j.tics.2011.11.005

Parat, C. (2013). L'affect p1artagé. *Rev. Fr. Psychosom*. 44, 167–182.

Pikler, E. (1962). *Que sait Faire Votre Bébé?*. Paris: Les éditeurs français réunis.

Rabeyron, T. (2016). Les processus de symbolisation et de représentation comme espace transitionnel pour la psychanalyse et les neurosciences. *Evol. Psychiatr.* 81, 160–175. doi: 10.1016/j.evopsy.2015.03.003

Rabeyron, T. (2017). Médiations thérapeutiques et processus de symbolisation : de l'expérience sensible à la modélisation. *Evol. Psychiatr.* 82, 351–364. doi: 10.1016/j.evopsy.2017.01.001

Rabeyron, T. (2018). Constructions finies et constructions infinies: de l'épistémologie psychanalytique dans ses rapports à la vérité. *Analysis* 2, 143–155. doi: 10.1016/j.inan.2018.07.002

Rabeyron, T. (2019). Processus transformationnels et champ analytique: un nouveau paradigme pour les modèles et les pratiques cliniques. *Evol. Psychiatr.* (in press).

Rabeyron, T., Chouvier, B., and Le Maléfan, P. (2010). Clinique des expériences exceptionnelles: du trauma à la solution paranormale. *Evol. Psychiatr.* 75, 633–653. doi: 10.1016/j.evopsy.2010.09.004

Rabeyron, T., and Loose, T. (2015). Anomalous experiences, trauma, and symbolization processes at the frontiers between psychoanalysis and cognitive neurosciences. *Front. Psychol.* 6:1926. doi: 10.3389/fpsyg.2015.01926

Raichle, M. E., and Snyder, A. Z. (2007). A default mode of brain function: a brief history of an evolving idea. *Neuroimage* 37, 1083–1090. doi: 10.1016/j.neuroimage.2007.02.041

Ramstead, M. J., Kirchhoff, M. D., and Friston, K. J. (2019). A tale of two densities: Active inference is enactive inference. *Adap. Behav.* 1–15.

Reik, T. (1936). *Surprise and the Psycho-Analyst: On the Conjecture and Comprehension of Unconscious Processes, 2014 Edn*. New York: Routledge.

Roussillon, R. (1992). *Du Baquet de Mesmer au Baquet de Sigmund Freud*. Paris: PUF.

Roussillon, R. (2001). *Le Plaisir et la Répétition: Théorie du Processus Psychique*. Paris: Dunod.

Roussillon, R. (2009). L'associativité. *Libr. Cah. Psychanal.* 20, 19–35.

Roussillon, R. (2011). *La Disposition D'esprit Clinique. In: Manuel de Pratique Clinique*. Paris: Elsevier.

Roussillon, R. (2012). L'associativité polymorphique et les extensions de la psychanalyse. *Carnet. Psy.* 162, 27–31.

Roussillon, R. (2015). An introduction to the work on primary symbolization. *Int. J. Psychoanal.* 96, 583–594. doi: 10.1111/1745-8315.12347

Roussillon, R., Chabert, C., Ciccone, A., and Ferrant, A. (2007). *Manuel de Psychologie et Psychopathologie Clinique Générale*. Paris: Masson.

Scarfone, D. (2018). Free association, surprise, trauma, and transference. *Psychoanal. Inquiry* 38, 468–477. doi: 10.1080/07351690.2018.1480232

Segal, H. (1957). Note on symbol formation. *Int. J. Psychoanal.* 38, 395–401.

Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. *Am. Psychol.* 65, 98–109. doi: 10.1037/a0018378

Solms, M. (2013). The conscious id. *Neuropsychoa* 15, 5–19. doi: 10.1080/15294145.2013.10773711

Solms, M. (2018). The hard problem of consciousness and the free energy principle. *Front. Psychol.* 9:2714. doi: 10.3389/fpsyg.2018.02714

Solms, M., and Panksepp, J. (2012). The "Id" knows more than the "Ego" admits: Neuropsychoanalytic and primal consciousness perspectives on the interface between affective and cognitive neuroscience. *Brain Sci.* 2, 147–175. doi: 10.3390/brainsci2020147

Solms, M., and Turnbull, O. (2011). What is neuropsychoanalysis? *Neuropsychoa* 13, 133–145. doi: 10.1080/15294145.2011.10773670

Steinert, C., Munder, T., Rabung, S., Hoyer, J., and Leichsenring, F. (2017). Psychodynamic therapy: as efficacious as other empirically supported treatments? A meta-analysis testing equivalence of outcomes. *Am. J. Psychiatr.* 174, 943–953. doi: 10.1176/appi.ajp.2017.17010057

Stern, A. (2011). Investigation psychanalytique sur le groupe borderline des névroses. Quelle thérapie engager?. *Rev. Fr. Psychanal.* 75, 331–348.

Stern, D. N. (2000). *The Interpersonal World of the Infant: A View From Psychoanalysis and Developmental Psychology*. New York, NY: Basic Books.

Stern, D. N. (2004). *The Present Moment in Psychotherapy and Everyday Life*. New York, NY: Norton.

Trevarthen, C., and Aitken, K. (2001). Infant intersubjectivity: research, theory, and clinical application. *J. Child Psychol. Psychiatr.* 42, 3–48. doi: 10.1111/1469-7610.00701

Ward, J. (2013). Synesthesia. *Annu. Rev. Psychol.* 64, 49–75. doi: 10.1146/annurev-psych-113011-143840

Widlöcher, D. (1996). *Les Nouvelles Cartes de la Psychanalyse*. Paris: Odile Jacob.

Widlöcher, D. (2010). Distinguishing psychoanalysis from psychotherapy. *Int. J. Psychoanal.* 91, 45–50. doi: 10.1111/j.1745-8315.2009.00233.x

Winnicott, D. W. (1958). *Through Paediatrics to Psycho-analysis*. London: Karnac.

Winnicott, D. W. (1963). *Fear of breakdown. Psychoanalytic Explorations (1974 Edn)*. Boston: Harvard UP.

Woll, C. F. J., and Schönbrodt, F. D. (2019). A series of meta-analytic tests of the efficacy of long-term psychoanalytic psychotherapy. *Eur. Psychol.* 25, 51–72.

Yovell, Y., Solms, M., and Fotopoulou, A. (2015). The case for neuropsychoanalysis: Why a dialogue with neuroscience is necessary but not sufficient for psychoanalysis. *Int. J. Psychoanal.* 96, 1745–1553. doi: 10.1111/1745-8315.12332

Conflict of Interest: 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 Psychology

HYPOTHESIS AND THEORY
published: 21 May 2020
doi: 10.3389/fpsyg.2020.00984

Unconscious Emotion and Free-Energy: A Philosophical and Neuroscientific Exploration

*Michael T. Michael**

Underwood International College, Yonsei University, Seoul, South Korea

Unconscious emotions are of central importance to psychoanalysis. They do, however, raise conceptual problems. The most pertinent concerns the intuition, shared by Freud, that consciousness is essential to emotion, which makes the idea of unconscious emotion seem paradoxical. In this paper, I address this paradox from the perspective of the philosopher R. C. Roberts' account of emotions as concern-based construals. I provide an interpretation of this account in the context of affective neuroscience and explore the form of Freudian repression that emotions may be subject to under such an interpretation. This exploration draws on evidence from research on alexithymia and utilises ideas from free-energy neuroscience. The free-energy framework, moreover, facilitates an account of repression that avoids the homunculus objection and coheres with recent work on hysteria.

Keywords: unconscious emotion, free-energy, psychoanalysis, repression, alexithymia, hysteria, construal

OPEN ACCESS

Edited by:
Peter Fonagy,
University College London,
United Kingdom

Reviewed by:
Mark Leonard Solms,
University of Cape Town, South Africa
Idit Shalev,
Ariel University, Israel

***Correspondence:**
Michael T. Michael
mmichael@yonsei.ac.kr;
mmichael.esq@gmail.com

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *24 December 2019*
Accepted: *20 April 2020*
Published: *21 May 2020*

Citation:
Michael MT (2020) Unconscious Emotion and Free-Energy: A Philosophical and Neuroscientific Exploration. Front. Psychol. 11:984. doi: 10.3389/fpsyg.2020.00984

INTRODUCTION

Freud appears ambivalent about emotion. On the one hand, he thought that it is "of the essence of an emotion that we should be aware of it, i.e., that it should become known to consciousness" (Freud, 1915/1957, p. 177). On the other hand, he frequently invoked unconscious emotion, such as "unconscious love, hate, anger, etc" (Freud, 1915/1957, p. 177). This ambivalence is reflected in an underdeveloped understanding of unconscious emotion in psychoanalysis today (Akhtar, 2013). The paradox suggested by Freud's apparently conflicting stances has not yet been fully resolved. As such, a primary aim of this paper is to address this puzzle and provide an account that makes sense of both the reality of unconscious emotion and the intuition that consciousness is essential to emotion.

The topic of unconscious emotion is no mere side-issue to psychoanalysis. It reaches into the foundations of the psychoanalytic enterprise. An important reason for this is the role that unconscious emotion plays in psychopathology. In a recently published paper (Michael, 2018b; building on Edwards et al., 2012), I argued that unconscious emotion may lie at the core of hysterical symptoms. The argument, in brief, is that the repression of a memory can lead to the repression of accompanying emotion. As a result, when the memory, and hence the accompanying emotion, is unconsciously triggered, the patient may experience the bodily feelings generated by the emotion, but these feelings lack any explanation due to the unconsciousness of the emotion. Unexplained bodily feelings constitute prediction error – or free-energy – according to the Bayesian brain framework (see section Free-Energy and the Process of Repression). Without the availability of the correct explanation for these feelings, the brain attempts to construct a plausible alternative explanation, which in the right circumstances would be a symptom "belief." This, in turn, can lead to the generation of symptoms of hysteria. If this account is correct, then it demonstrates the important role that unconscious emotion plays in the emergence of the kind of phenomena that psychoanalysis was first designed to address.

Given this, a broader aim of this paper is to address the Freudian paradox about unconscious emotion in a way that also sheds additional light on psychopathology. To do so I will invoke a philosophical account of emotion. This relates to a secondary aim of the paper, which is to bring a philosophical perspective into dialogue with psychoanalytic and neuroscientific perspectives. I believe that doing so, though challenging, is necessary to providing a more rounded and comprehensive understanding of the present issues. Emotion, though addressed by psychoanalysis and neuroscience, is not a concept that derives from these disciplines, but rather from our everyday psychological discourse, and philosophers have spent the last few decades analysing just such concepts. As such, the philosophy of emotion offers a prism through which a subtler understand of unconscious emotion may be attained.

The focus of the paper, to be more specific, will be on the *repression of the consciousness of emotion*. I use this cumbersome phrase to hone in on the form of repression at stake. Freud, despite his aforementioned comments, did speak about the repression of emotions, except that what he actually spoke about in this context was (chiefly) the *suppression of affect*. He wrote, "to suppress the development of affect is the true aim of repression and… its work is incomplete if this aim is not achieved" (Freud, 1915/1957, p. 178). There is a distinction, for Freud, between the suppression of affect and the repression of ideas, which is that "unconscious ideas continue to exist after repression as actual structures in the system *Ucs*., whereas all that corresponds in that system to unconscious affects is a potential beginning which is prevented from developing" (Freud, 1915/1957, p. 178). Thus, (full) suppression of affect cannot co-occur with (full) emotion, since on Freud's account such suppression prevents the development of the emotion. What I wish to focus on instead, and which Freud, strictly, denied as a possibility, is the case where the emotion occurs – indeed, with felt consequences, as in the unexplained bodily feelings that, on my account, hysterics interpret as due to a symptom – but where *consciousness of* this emotion is prevented from arising due to repression.

In all probability, there are many gradations in the repression of emotion. Psychological defence against emotion, in other words, may bring about effects that fluctuate between numerous levels. These levels plausibly include the following: (1) suppression of the behavioural expression of an emotion that the agent is nevertheless acutely aware of; (2) repression of the consciousness of the emotion, as discussed above; and (3) full suppression of the emotion. As stated, our focus will be on the second level, since it is this which most relates to hysteria, and is the key to understanding Freud's seemingly paradoxical comments on unconscious emotion.

A CONSTRUAL ACCOUNT OF EMOTIONS

We begin with a philosophical account of emotion. The philosophy of emotion is important insofar as emotion is a concept that derives from commonsense psychological discourse, that is, the everyday discourse by which we make sense of our own and other people's behaviour in terms of mental states such as belief, desire, and emotion. As such, it seems sensible to begin an endeavour to understand emotion by observing the constraints that our commonsense discourse sets on the concept. This is a motive for engaging in what analytic philosophers call "conceptual analysis," the attempt to analyse concepts according to our most basic intuitions about their use in everyday language.

In order to motivate the account of emotion I will be presenting in this paper, I will first briefly offer some historical context. It is beyond the scope of the present paper to give anything other than a cursory review, but I will mention a couple of relevant developments within the recent history of the philosophy of emotions. One theory prominent in the late nineteenth and early twentieth century was the feeling theory, asserting that emotions are conscious feelings. A classic example of a feeling theory is the James-Lange theory (James, 1884; Lange, 1885), which posits that emotions are the perceptions of physiological changes in the body. Scientists and philosophers, however, soon observed that there were a number of problems faced by such theories. These include problems with accounting for the differences between emotions (since the feeling profiles of different emotions are often remarkably similar, while the feeling profiles of instances of the same emotion can differ widely), accounting for the rational dimension of emotions (drawing on the observation that emotions are subject to justification), accounting for the intentionality of emotions (in the sense of their being *about* some object), and accounting for the strong association between emotions and evaluations (e.g., fear seems to correspond in some way to evaluating an object as dangerous) (Scarantino and de Souza, 2018). Such problems do not entail that feelings theories should be dismissed, but they do require that such theories should be sophisticated enough to address these issues.

Motivated by the desire to deal with the problems brought against feeling theories, many philosophers moved in a different direction, developing judgement theories of emotion. Such theories assert that emotions are judgements (e.g., Neu, 1977; Solomon, 1980; Nussbaum, 1990). Fear, for instance, is the judgement that some object poses a danger to oneself. While such theories were popular for a while, they too encounter problems. An important issue is accounting for irrational emotions (Stocker and Hegeman, 1992). For example, an arachnophobe may judge a spider to be of no threat to him whatsoever, yet still fear it, suggesting there is a gap between judgement and emotion.

This brings us to the theory that I will be discussing in this paper: construal theory. The emergence of construal theories is a more recent development in the philosophy of emotions, arising as a result of criticisms of judgement theories. They offer an advance on judgement theories in that they can account for irrational emotions while still providing convincing solutions to the problems faced by feeling theories. In this paper, I will focus on a particular construal account that has been influential in the philosophical literature on emotion and provides a relatively simple yet

plausible account of emotions. There are other versions of construal theory, but they have many of the same features as this one (see Lacewing, 2004, for a review). The account, proposed by Roberts (1988, 2003), is that emotions are concern-based construals.

In order to understand this account, we need to understand what Roberts intends by "construal." The concept was inspired by a passage from the *Philosophical Investigations* (Wittgenstein, 1953, pp. 193–194), in which Wittgenstein talks about a sense of "sees" that is different from that of bare perception. He illustrates this sense with the famous duck-rabbit illusion: one can *see* the figure *as* a duck or *as* a rabbit. The seeing of some object, X, as something else, Y, is what Roberts calls a construal.

It is important to note that, though the concept of construal was inspired and is best illustrated by perceptual examples, it is not limited to such. As Roberts (1988, pp. 190–191) states, the elements of a construal, the X and Y terms, can be various – for example, they can be percepts, thoughts, images, concepts, or combinations of these. Elsewhere (Michael, 2018a, 2019b), I have argued that construal need not be conceptual in character – that is, the Y element need not involve concepts. Thus, for example, an infant may see a stranger as threatening even though she does not have the concept of threat. She sees the stranger through a set of experiences, involving perhaps various feelings, memories, imaginings, and so on, such that those experiences colour her experience of the stranger in a certain way, where this way is appropriately described as the aspect of being threatening.

Following Wittgenstein, Roberts takes a "family resemblance" approach to concepts, thus he does not hold that "construal" (or "emotion" for that matter) can be captured by a set of necessary or sufficient conditions. I agree, but nevertheless it will be helpful to adopt at least a working definition. To this end, and taking into account the point made above that construal need not be conceptual, I define construal as a way in which an intentional agent experiences or responds to some object, X, where this way of experiencing or responding can be appropriately described by phrases of the form "as Y" or "in terms of Y."

According to Roberts, emotions are construals. A good way of understanding this is via Richard Lazarus's influential appraisal theory of emotion. For Lazarus (1991), emotions involve appraisals, in the form of evaluations that he calls "core relational themes." For example, anger involves appraising some object as having caused "a demeaning offence against me or mine"; fright involves "facing an immediate, concrete, and overwhelming physical danger"; and disgust involves "taking in or being too close to an indigestible object or idea" (Lazarus, 1991, p. 122). Lazarus holds that we make such appraisals in the form of judgements. On Roberts' account, however, emotions need not involve judgements. Instead they are construals in the form of "see X as Y" where Y corresponds to a core relational theme. Thus, to be angry with a person is to see that person in terms of "a demeaning offence against me or mine" – that is, according to my definition, to experience or respond to the person in a particular way, where that way is appropriately described in terms of the given evaluation. To offer another example, the infant's fear of the stranger is constituted (in part) by her seeing the stranger as threatening, that is, by her relating to him via a set of experiences or responses that can be collectively described as the aspect of being threatening[1].

The other part of what constitutes an emotion, on Roberts' account, is concern. By this he means a range of states which we can broadly term "desires" (relating to approach behaviours) or "aversions" (relating to avoidance behaviours)[2]. Emotions are concern-based construals, that is, construals filtered through desires or aversions. Such concerns enter into the construal as part of the Y-term: when I am angry with someone, I am not just seeing them in terms of having culpably offended against me, but also in terms of my concern not to be so offended against. Similarly, the infant is relating to the stranger via her aversion to threatening objects. Thus, an emotion is not simply an evaluative construal, but is rather one in which a concern is interwoven with the evaluation.

Among the merits of this account of emotion is that it makes sense of the explanatory role that emotions play in commonsense psychological discourse. Emotions typically explain subsequent behaviours and are explained by preceding events. We can explain my aggressive behaviour towards the person who has angered me, for example, by my concerned construal of the person as having offended against me, and we can explain this state in turn in terms of that person behaving towards me in a way that can plausibly be construed as offensive. The advantage that Roberts' account has over judgement theories of emotion is that there are many cases in which one may have an emotion despite also having a judgement contrary to the evaluation associated with that emotion. These are, as we have seen, the "irrational emotions," such as phobias. The arachnophobe may well judge that the spider before him is harmless, yet nevertheless be afraid of it. On Robert's account, he is construing the spider as threatening, while judging it not to be. Elsewhere, I have elaborated on this distinction between judgement and construal, and illustrated how it can help solve numerous philosophical problems (Michael, 2018a).

In order to understand the power of this construal theory, it would be instructive to compare it with a recent feeling theory of emotion. The philosopher Prinz (2004) has offered a compelling update to the James-Lange theory of emotions, one that is sophisticated enough to address the problems brought against simple feeling theories. On this account, as in the James-Lange theory, the perception of bodily changes is constitutive of emotion. But, Prinz argues, this does not mean that we should give up on the idea that emotions are essentially evaluative. Rather, the perception of bodily changes can itself represent a core relational theme, so that emotions can be evaluative without being conceptual[3]. In other words, emotions are embodied

[1]On this account, where having an emotion is to construe X as Y, the X is what philosophers call the "particular object" of the emotion, and the Y is the "formal object" (Scarantino and de Souza, 2018).

[2]This "concern" dimension of emotion resonates with Freud's idea that an emotion is a manifestation of a drive (*Trieb*), since it accounts for the directedness of an emotion towards some need-based outcome.

[3]This relies on Prinz's theory of representation (Prinz, 2004, pp. 52–78), for which it suffices that a state is reliably caused by something (e.g., a core relational theme) and has been set up (by learning or evolution) to be reliably caused by that thing.

appraisals: "They represent core relational themes, but they do so by perceiving bodily changes" (p. 68).

Prinz does not regard his theory as a construal theory, as he thinks that such theories assume the "conceptualisation" and "disembodiment" hypotheses. The conceptualisation hypothesis is the claim that "emotions require concepts" (p. 23). The disembodiment hypothesis is the claim that the components of emotion are "not identical to bodily changes or internal states that register bodily changes" (ibid.). I believe, however, he is wrong about the assumptions of construal theories. As we have seen, the version of construal theory I have presented does not assume the conceptualisation hypothesis, since the elements of a construal need not involve concepts. Also, it does not imply the disembodiment hypothesis: it allows that one can construe X as Y in virtue of a perception of bodily changes. Hence Prinz's theory can be seen as a construal theory. Consider the example of one who feels fear upon seeing a snake. According to Prinz's theory, the perception of the bodily changes brought on by the sight of the snake represents the core relational theme of danger, and it is this perception which constitutes the emotion of fear. On my interpretation of construal, this is the same as claiming that one construes the snake as dangerous, where this construal is in virtue of a perception of bodily changes[4].

I offer this argument not as a means of endorsing Prinz's theory, but to illustrate how a construal theory, broadly conceived, can accommodate sophisticated feeling theories, that is, ones that take into account the evaluative aspect of emotions. Construal, as I have defined it, is broad enough to encompass embodied non-conceptual construals. As such, it would be a mistake to describe construal theory, as some do (e.g., Smith and Lane, 2015), as a cognitivist theory. Though construals can be cognitive, they need not be. While resembling cognitions in respect of representing evaluations, they can consist solely of conscious feelings. This ideally suits them to an account of emotion, since, as we have seen, emotion has been analysed by some as a cognition and by others as a feeling, thus defining emotion in terms of construal allows for a compromise between these two positions.

Where Roberts' account becomes most useful for our purpose of understanding unconscious emotion is in relation to the question of how we feel emotions. The word "feeling" can mean different things, but an important sense of the word is, according to Roberts (1988), captured by construal. This sense is different from that of the bodily feelings that may constitute the emotion. It is, rather, what Roberts calls a feeling of construed condition, that is, of taking oneself "to be in a certain condition" or "to have a certain property" (p. 185). For example, to feel excluded is to take oneself as being excluded (p. 186). It is this sense of feeling that, according to Roberts, is most relevant to the locution "feeling an emotion." Thus, feeling our emotion involves a construal of our construal. Here is how Roberts (2003, p. 320) explains this idea:

> Let us use subscripts to distinguish the two construals, a subscript 1 for the emotion and a subscript 2 for the feeling, and place brackets around the word "construal" to indicate that the ordinary subject does not experience his emotion in terms of the concept of a construal. Thus, to feel angry at Sally is to construe_2 oneself as [construing_1] Sally as having culpably offended in some matter that one strongly cares about. To feel proud of Nathan is to construe_2 myself as [construing_1] myself as increased in status because of Nathan's attributes. To feel contrite is to construe_2 myself as [construing_1] myself as being or having done something contrary to some moral or quasi-moral standard that I am strongly concerned to meet.

So the feeling of an emotion is a (conscious) second-order construal, where what is construed is *oneself* in terms of a first-order construal. That is, when one feels an emotion one sees *oneself* in terms of *the way one is experiencing or responding to* some *object*[5]. This is a relationship between three elements: oneself, an object, and the way one is experiencing or responding to that object.

I will adopt Roberts' account of feeling an emotion, though I prefer to call it the *consciousness of an emotion*. By adopting this terminology, I am not thereby implying that, in the absence of a second-order construal, an emotion cannot, in some sense, be conscious. Supposing, as we entertained earlier, that the emotion is constituted by the experience of bodily changes. Then the emotion is conscious even in the absence of a second-order construal insofar as those experiences are conscious. However, the person who has the emotion need not be conscious of it *as* her emotion, that is, as her taking some object in terms of a particular evaluation. She need not, in other words, be *reflectively* conscious of her emotion. It is this subtle distinction between different senses of consciousness that will help us address the Freudian paradox.

A NEUROSCIENTIFIC INTERPRETATION

Let me pause here to explore how these ideas might translate into terms more familiar to neuroscientists. In doing so, I caution against seeing the forthcoming discussion as an attempt to *reduce* Roberts' account to neuroscientific terms. Roberts' is a philosophical account, that is, an attempt to analyse the concept of emotion as it occurs in its "natural habitat" of commonsense psychological discourse. There is no good reason to think that either the concept of emotion or the concepts used to analyse emotion, such as construal, should correspond neatly to neuroscientific concepts. To borrow another idea from Wittgenstein, commonsense psychology and neuroscience may be different "language games" that cannot be fully reconciled. Nevertheless, I do not go as far as some philosophers in considering the two domains to be entirely autonomous. I believe, rather, that it is reasonable to expect a rough correspondence between what happens in the brain and what happens "in the mind," so to speak, hence we can aspire to an approximate relation of ideas, as I hope the forthcoming discussion will illustrate.

[4]Here I am reproducing an argument I made in Michael (2019b).

[5]The object of the emotion is whatever object the emotion is about. This is typically a particular (possibly non-existent) entity, though can also be a state of affairs.

Our focus will be on typical emotional episodes related to the basic emotions identified by Jaak Panksepp. Panksepp (1998) uses the term "emotional command system" to designate brain systems that, upon certain input, "generate instinctual behaviour output patterns" (p. 28) that can be associated with common emotions (or related states). He specifically identifies seven such systems, of FEAR, RAGE, LUST, CARE, PANIC/GRIEF, SEEKING, and PLAY. These systems form the basic, innately programmed response to relevant stimuli, though what makes a stimulus relevant and the precise nature of the response require individual learning. As Solms (2019, p. 8) puts it, "fear behaviours (freezing and fleeing), for example, are innate predictions; but each individual has to learn what to fear and what else might be done in response." My purpose in drawing a connection with these ideas is to show how Roberts' account of emotions may relate to neuroscientific accounts, though the sketch I present is a simplified one.

The starting point of a typical emotional episode, on this account, is a "stimulus" that "triggers" an emotional reaction[6]. This stimulus may be (the perception of) an external event or may be an internal event, such as a thought or memory. It may or may not involve a cognitive interpretation of the event (i.e., a judgement or cognitive construal). The proposal is that this stimulus triggers a basic emotional command system, thereby setting in motion physiological changes preparing the body for a particular kind of behavioural response – possibly alongside cognitive changes, such as changes in the "style and level of efficiency of cognitive process" (Damasio, 1994, p. 163) – pertinent to the basic emotion triggered. Such physiological changes may include changes in heart rate, blood pressure, breathing, metabolism, release of hormones, and so on. These physiological changes are experienced by the agent through conscious affective feelings. The feelings are, in the main, ones of valenced (i.e., pleasant or unpleasant) arousal (Barrett, 2017, p. 72), though the combination of effects generated by the stimulus may have numerous distinctive features[7].

At this point, I will present one attractive possibility for interpreting Roberts' account of emotion, though later I will challenge this interpretation. Continuing on from the above description, we note that the agent, in perceiving or thinking about the stimulus, will do so via the affective feelings generated by it, so that the stimulus is experienced in a particular way. It is therefore tempting to see these feelings (plus related memories, fantasies, beliefs, and so on) as thereby constituting the "colouring" with which some object (e.g., a person) is experienced. Because this "colouring" is valenced and related to specific behavioural tendencies, it serves as a particular perspective on or evaluation of the object. For example, the unpleasant arousal aiming at a flight or freeze reaction generated by the FEAR system in response to the perception of a snake constitutes (in part) the evaluative aspect of seeing something as threatening. In other words, the feelings accompanying inner bodily changes, through which the object that triggered this reaction is now being experienced, represent an evaluation of that object. Experiencing an object in terms of such feelings is, as we have already noted, an *embodied construal*: as a consequence (largely) of bodily changes, one is experiencing the object, X, in a certain way, where that way can be appropriately described using an "as Y" phrase with Y being an evaluation of X. Thus, we may conclude that the construal that constitutes an emotion (on Roberts' account) is (usually) just such an embodied construal.

This interpretation has some nice features. First, it shows how a construal account captures the essential characteristics of emotion. The embodied construal described above is part of a causal chain that explains subsequent behaviour, and, when supplemented with an understanding of the predisposing tendencies of different kinds of stimuli, can be explained by preceding events. An embodied construal of this kind has the intentional (in the sense of being about some object) and evaluative character of emotion, in that it represents the evaluation of an object. It also has its motivational character, in that, in being a concern-based construal, it can predispose the agent towards certain actions (inherent in the behavioural preparedness triggered by the stimulus). Second, as we have already seen, the above interpretation also serves as a compromise between competing prominent theories of emotion, such as the James-Lange theory and appraisal theories, by utilising the idea that the experience of bodily changes is part of what constitutes an embodied appraisal. Third, in being a manifestation of a construal account, such an account is not tied to the basic or typical cases of emotional episodes described above, but potentially has wider applicability. Fourth, and most important for our purposes, this interpretation of emotion also captures a sense in which Freud's assertion that emotions are always conscious might be true, for the affective feelings that are essential to how the object of the emotion is construed are conscious experiences.

However, though I am tempted by such an interpretation, I think it is not quite right. This is not to say it is completely false: experiencing an object via a set of conscious affective feelings is a construal, and it is part of the construal that constitutes an emotion. But it is not, or need not be, the whole of it. The construal which constitutes the emotion refers, rather, more wholly to *the organism's response* to an object, beginning with the triggering of the emotional command system, up to and including the experiencing of the object via the conscious affects and other effects of this triggering. This entire organismic response is, I believe, what constitutes the construal that defines emotion, since the response as a whole (and not just some part of it)

[6]The terms "stimulus" and "triggers" are perhaps misleading, as perceptions, thoughts, emotions, and so on, are, on a free-energy account, the result of simultaneous cascading predictions (Barrett, 2017). Nevertheless, the terms are expositorily convenient, and, since our focus is not on the initiation of emotion but its subsequent course, I will continue to use them in the forthcoming discussion.

[7]Here, in order not to complicate matters, I skate over the distinction between what Prinz calls the "perception of bodily changes" and the affect, that is, the conscious feeling on which this perception is based. Solms and Friston (2018) argue that affect is the subjective manifestation of forebrain arousal by the brainstem, triggered by prediction error. I suggest that this may relate to Prinz's "perception of bodily changes" through a construal: I feel X (some body region) in terms of Y (affect). If this is correct, then the "perception of bodily change" is itself a construal, but different from the construal that is the emotion (though this may construe *in virtue* of such perceptions).

can be taken as the construed aspect[8]. Take for example the case of the perception of a snake triggering a fear response in an organism. The organism's entire response to the snake, from the initial triggering of the emotional command system through to the experiencing of the snake via the arousal and other concomitant effects generated by the command system, constitutes that organism's evaluative construal of the snake. Responding to an object in this way is as much an embodied construal as experiencing the snake in terms of the affective feelings generated: one is responding to an object, X (e.g., the snake), in a certain way, where that way can be appropriately described using an "as Y" phrase ("as threatening"), with Y being an evaluation of X[9].

Henceforth I will refer to the first possibility, in which a construal and hence an emotion are constituted by the way an agent experiences an object, as a *narrow account* of construal and emotion, and the second possibility, in which a construal and hence an emotion are constituted by an organism's response to an object, as a *broad account* of construal and emotion[10].

Adopting a broader account of emotion renders the question of unconscious emotion less problematic. For on a narrower construal account, which focuses on how one experiences a certain object, consciousness is essential to the emotion. But on the broader account, one may be having an emotion without this necessarily having an effect on one's conscious experience. In practice, this may make little difference, as an emotional episode will almost always influence how one experiences an object, but the conceptual distinction does at least allow for the possibility of entirely unconscious emotions. It seems to me, therefore, that the paradox of unconscious emotion, which Freud himself touched on, may arise as a result of adopting a narrower conception of emotion than is required. Nevertheless, as we proceed, it is worth having both accounts in mind, as the first account, even if incorrect, will help as understand why many, like Freud, have seen the idea of unconscious emotion as paradoxical.

LEVELS OF EMOTIONAL AWARENESS

It would be useful to connect the above ideas with an influential neuroscientific model of emotional consciousness, which I call the *Levels of Emotional Awareness* (*LEA*) model[11]. This model, inspired by Marr's (1982) three-level theory of vision, has been most clearly articulated by Prinz (2004) and Lane et al. (2015). It posits that emotional consciousness is based on three levels of processing. The lowest level of the hierarchy pertains to local bodily states, that is, for example, changes in visceral states, changes in hormonal levels, and so on (Prinz, 2004, p. 213). Anatomically, Lane et al. (2015, p. 603) associate this level with the activity of brainstem nuclei. The intermediate level involves integrating these first-level processes into coherent patterns, ultimately "patterns of one's entire bodily state across organs, muscles, and so forth" (p. 599). Anatomically, according to Lane et al. (2015), this level corresponds most closely with activity in the insula, "a predominantly sensory structure that registers and remaps bodily information and sensations into conscious somatic sensations" (p. 602)[12]. The highest level involves abstracting from particular patterns by categorising a range of such patterns under the same representation, that is, as "having the same emotional meaning" (ibid.). Lane et al. (2015) argue that this level of processing is associated with activity in the rostral anterior cingulate cortex (rACC), a region of the brain which specialises "in the representation of emotional meaning, particularly meaning that is concept-driven, by integrating highly processed interoceptive and exteroceptive information" (ibid.).

The LEA model may be useful in anchoring some of the ideas presented in the previous section. The first and second levels of processing described above, associated with activity in the brainstem and insula, correspond most closely with the affect and experience of bodily changes accompanying (and perhaps partly constituting) an emotion[13]. More importantly for our purposes, the highest level of processing in the LEA model, associated with the activity of the rACC, corresponds most closely with the second-order construal that constitutes the feeling of an emotion on Roberts' account, for it is at this level that meaning is assigned to the emotional episode. As Lane et al. (2015) explain, "if the high-level of body state representation malfunctions then one will still experience and respond to bodily states, and other people will recognise them as expressions of emotion, but one will not experience them *as emotions*, be able to label them as such, or be able to use knowledge of their emotional meaning to plan to respond to them appropriately" (p. 599; authors' emphasis).

Further support for the correspondence between second-order construal and the highest level of processing in the LEA model comes from Stevens (2016), who describes several lines of evidence suggesting that the consciousness of emotion is closely associated with rACC activity. For example, "studies examining the rACC region in alexithymia [a condition of reduced emotional awareness; see section Evidence From Alexithymia] show a pattern of hypoactivation" (p. 58). Also, in studies of different subtypes of depression, "a pattern emerges in which

[8]This is in line with a free-energy account, since the action plan generated by the emotional command system may be seen simultaneously as an inference about the causes of sensory input. That is to say, as the perception of the "stimulus" is being constructed, the brain is at the same time predicting the body's needs in response to this stimulus, hence in preparing a particular action plan (e.g., for a "flight" response) it is also thereby evaluating the stimulus (say, as threatening). The action plan and the evaluation are not two distinct states, but rather are two sides of the same coin.

[9]The response meets Prinz's (2004) criteria for a representation (see footnote 3), since it is reliably caused by a core relation theme (such as danger) and has been set up by evolution and learning to be so reliably caused.

[10]The narrow account of construal fits best with the phrase "sees X as Y," while the broad account of construal fits best with the phrase "takes X as Y."

[11]The name derives from Lane et al.'s (1990) "Levels of Emotional Awareness Scale."

[12]It should be noted that the role of the insula in emotional awareness is contested (Damasio et al., 2013). While it is widely believed that the insula normally plays a role in emotional awareness, it is not yet known precisely what this role is, and it may be that other brain areas can perform similar functions in cases where the insula has been damaged (ibid., p. 844).

[13]Solms (2013, 2019) and Solms and Friston (2018) make a powerful case that affect is generated by the brainstem, while perceptions relate to cortical regions. This seems consistent with Prinz's (2004) account of the perception of bodily changes, which he associates with the insula (p. 215). Thus, on this view, the first level of the LEA corresponds to affect, and the second level to the perception of bodily changes in terms of such affect (see footnote 7).

those that have awareness of their feelings show hyper rACC activation and those that are unaware of their feelings show hypo rACC activity" (p. 59).

Lane et al. (2015) also bring to attention another important dimension of the consciousness of an emotion, which is that it involves "situational appraisal." They associate such appraisal with the ventromedial prefrontal cortex (vmPFC), stating that "one can think of this area as participating in the ongoing evaluation of emotional significance of stimuli in the environment in communication with cortical structures such as the insula and subcortical structures such as the amygdala, and generating representations of the emotional meaning of one's situation" (p. 602). This kind of appraisal seems pertinent to second-order construal, since such is concerned with representing the meaning derived from one's affective response to elements in the environment (i.e., one's first-order construal of those elements). Focusing on such situational appraisal also brings to the fore the importance of context to the consciousness of emotion. The nature of an emotion cannot simply be read off the affective feelings it generates – indeed there may be no accurate mapping from quality of affective feeling to emotion (Barrett, 2017, p. 112). Rather an emotion needs to be understood in relation to a situational context, for, on the construal view of emotion, the emotion is an evaluation of some stimulus, where the nature of that evaluation depends on the wider circumstances in which that stimulus arose (Eickers et al., 2017). As we will see in section Free-Energy and the Process of Repression, this situational dimension can be important in determining why, in some cases, the emotion is repressed.

SOLUTION TO THE FREUDIAN PARADOX

As mentioned, the idea of unconscious emotion has been seen to present something of a paradox. This is because, in accord with Freud, many find it intuitive that consciousness is intrinsic to emotion. Yet this intuition has been challenged (e.g., Pulver, 1971) and the consensus within contemporary psychoanalysis is that emotion can be unconscious (Akhtar, 2013, pp. 14–15). Indeed, Freud himself acknowledged that talk of unconscious emotions is widespread in psychoanalysis:

> But in psycho-analytic practice we are accustomed to speak of unconscious love, hate, anger, etc., and find it impossible to avoid even the strange conjunction, "unconscious consciousness of guilt," or a paradoxical "unconscious anxiety" (Freud, 1915/1957, p. 177).

Freud does indeed make numerous references to unconscious emotion throughout his work (e.g., Freud, 1900/1957, p. 560, 1905/1957, pp. 56–57, 1909/1957, p. 240, 1910/1957, p. 144, 1911/1957, p. 63, 1919/1957, p. 231, 1933/1957, p. 139). So before we examine how repression works in relation to emotions, we need to first say more about this apparent paradox.

As I postulated in section A Neuroscientific Interpretation, one potential solution to the paradox is that Freud was adopting too narrow a view of emotion, one for which conscious experience is essential, whereas there is a broader view of emotion in which conscious experience is not essential. Hence emotion can be unconscious when taken in the broader sense, though is necessarily conscious when taken in the narrower sense.

But there is also another solution available, one that works even if we adopt only the narrow sense of emotion. This second solution to the paradox is suggested by Roberts' account of what it is to feel an emotion. Recall that, for Roberts, having an emotion is having a first-order construal, while feeling an emotion is having a second-order construal, that is, a construal of oneself as construing some object in a certain way. This allows us to distinguish between two forms of consciousness: the conscious experiences that (partly) constitute the emotion and the *consciousness of* the emotion. In relation to the narrow interpretation of Roberts' account, which focuses on embodied construal as a way of experiencing some object, this distinction can be stated as that between affective consciousness (feeling in the sense of affective feeling) and the consciousness of the emotion (feeling in the sense of feeling as a construed condition).

It is worthwhile saying a little more about the nature of the consciousness of an emotion. To do so we need to pay closer attention to the characteristics of the second-order construal (cf. Damasio, 1999). Whereas I have stated that a construal need not be conceptual, a second-order construal of the kind we are currently contemplating is conceptual. The experience that constitutes the emotion, itself an integration of various experiences of bodily change (and possibly of non-bodily changes), is construed as an instance of a particular kind of experience or response. Simultaneously, this is directly related to some object, so that it is a way of experiencing or responding to that object. At the same time this is understood as *one's* way of experiencing or responding to the object. Such an integration seems only achievable by relating these elements conceptually. In the simplest case, one comes to construe these felt changes as, say, one's anger at X, though such straightforward emotional labelling is not a requirement of the consciousness of an emotion[14], but rather what matters is that one has a coherent and articulable perspective on the object.

REPRESSION OF THE CONSCIOUSNESS OF EMOTION

The above ideas readily lend themselves to the following characterisation of emotional repression[15]. The repression of the consciousness of an emotion is an active process that seeks to reduce attention on – or the precision of one's model of (as we will see in section Free-Energy and the Process of Repression) – how one is experiencing or responding to the object of the emotion. This account of the repression of the consciousness of emotion has the advantage that it unproblematically allows that an agent

[14] Roberts (2003, p. 321) writes, "I do not suggest that we cannot feel an emotion (in my sense of "feel") unless we have a name for the emotion. The important thing is that we have ways of conceptualising ourselves, and I should think it obvious that we have a lot more concepts than we have concept words."

[15] In using the term "repression" I am following Freud's usage in *The Unconscious* (Freud, 1915/1957, p. 178).

can have an emotion, where that emotion is accompanied and partly constituted by conscious affective feelings, without being conscious of it, for the first-order construal that is the emotion need not be affected by the repression.

Such an account leads to some interesting reflections, which I will state in the form of a problem and suggested solutions. The problem is this: How, if an agent is experiencing the bodily changes involved in the emotion, can the repression of the second-order construal be sustained? For, it may be argued, the agent would surely need to interpret those experiences in some way.

There are at least two possible solutions to this problem. The first is that, though the correct second-order construal of the emotion is repressed, another, incorrect, construal can be constructed that offers an explanation of sorts for the given experiences. Indeed this possibility is, arguably, suggested by Freud:

> In the first place, it may happen that an affective or emotional impulse is perceived but misconstrued. Owing to the repression of its proper representative it has been forced to become connected with another idea, and is now regarded by consciousness as the manifestation of that idea. If we restore the true connection, we call the original affective impulse an "unconscious" one. Yet its affect was never unconscious; all that happened was that its *idea* had undergone repression (Freud, 1915/1957, pp. 177–178).

There has been some discussion among psychoanalytic scholars as to how best to understand what Freud means by "proper representative" (e.g., Green, 2004; Herrera, 2010). The dominant view is that such a "representative" is a mental representation of the object of the emotion (Boag, 2012, p. 33), so that what Freud is talking about above is merely a displacement from one object to another. But there is another possible interpretation – which even if not exegetically correct, may be more theoretically appropriate – in line with my account. This is that the "proper representative" of the emotion is the second-order construal that constitutes the consciousness of an emotion. Thus, we can interpret Freud's above assertion as that repression can cause an inaccurate second-order construal of one's emotion to arise. Such a second-order construal can be inaccurate by misrepresenting the object of the emotion, as the standard interpretation asserts (corresponding to seeing one's seeing X as Y as one's seeing *A* as Y); or by misrepresenting the emotion as a different emotion by associating it with a different set of evaluative concepts (seeing one's seeing X as Y as one's seeing X as *B*); or even by misrepresenting the subject of the emotion as other than the self, thereby constituting projection (seeing one's seeing X as Y as *S*'s seeing X as Y). Hence, in a more literal sense than that provided by the standard interpretation, an "emotional impulse is perceived but misconstrued."

A second and more important answer to the question of how the unconsciousness of an emotion can be sustained in light of the conscious affective feelings it generates is that this is, in many cases, precisely the problem that leads to pathology. By repressing the second-order construal, one is left with unexplained experiences that constitute the prediction error that drives neurotic symptoms, as postulated by my Freudian version of the Bayesian account of hysteria (Michael, 2018b), described in the introduction. To say a little more about this, consider an agent who has repressed the consciousness of her emotion. Especially if she has increased bodily awareness (perhaps due to trait interoceptive sensibility, or increased body focus due to illness), she is likely to experience the bodily changes generated by the unconscious emotion while being unable to explain them. In which case, the repression becomes a "force" that compels her brain-mind towards alternative explanations. These alternative explanations may include a symptom "belief"[16] (which can arise due to numerous factors, such as recent experiences with illness, cultural or other illness-related beliefs, or apt symbolic correspondences). As long as such a "belief" offers a plausible explanation, it may, due to the repressive need to keep attention away from the correct explanation, be favoured by Bayesian processes to the point where it becomes entrenched – that is, it is afforded a degree of precision the makes it immune to revision in the light of contrary sensory evidence. Such an entrenched symptom "belief" can thereby come to generate the symptom (see Michael, 2018b, for more details).

EVIDENCE FROM ALEXITHYMIA

The proposal that unconscious emotion involves the repression of a second-order construal of one's emotion has support from work on alexithymia. Alexithymia is a condition characterised by an inability to gain awareness of one's emotion and to express it in words[17]. It has often been cited in the philosophical literature on emotions as exemplifying unconscious emotion. For example, Lacewing (2007, p. 22) brings up alexithymia as "cases in which the subject reports no particular feelings at the time of the emotional episode," stating that:

> They generally disavow feeling emotions, and so they are also known as "alexithymics" (from the Greek for "having no words for emotion"). However, on the basis of how they interact with other people and the emotions they arouse in others, psychoanalysts argue that they do in fact have emotions, but that they are very out of touch with them.

The scientific literature on alexithymia suggests that, though alexithymics are not aware of their emotions (that is, according to Roberts' account, they do not feel their emotions), they do feel the bodily sensations associated with the emotions. As Liemburg et al. (2012) put it, "alexithymia is characterised by difficulty to distinguish emotions from bodily sensations" (p. 660), so it is by failing to distinguish emotions from bodily

[16]"Belief" here is not to be understood in the usual sense, as a propositional attitude. It is rather a representation encoded by the activity of a population of neurons, occurring as part of a hierarchical model of the causes of sensory input (see section Free-Energy and the Process of Repression).

[17]See Taylor and Bagby (2013) for a more in-depth understanding of the alexithymia construct, including its historical background and empirical grounding. For a psychoanalytic perspective on the condition, see McDougall (1982, 1989).

sensations, rather than not feeling those sensations, that the problem (in part) arises. Moreover, the same authors found evidence for "a diminished connectivity within the DMN (default mode network) of alexithymic participants, in brain areas (such as the ACC) that may also be involved in emotional awareness and self-referential processing" (ibid.) – that is, just the kind of pattern we might expect in relation to a second-order construal that integrates the self with representations of one's emotional state. These considerations cohere with the idea that the consciousness of an emotion is distinct from both having the emotion and from the consciousness of affect that may be partly constitutive of the emotion (at least, on a narrow account of emotion).

Interestingly, alexithymia has a high comorbidity with numerous psychiatric disorders:

> Alexithymia has been associated with increased risk for psychosomatic complaints, anxiety disorders and depression. and the emotion regulation difficulties characteristic of alexithymia have been hypothesized to play a mediating role in these (ibid.).

Of particular relevance is the comorbidity with psychosomatic complaints, which, as characteristic of hysteria (or conversion disorder), may be a prime example of the pathology of repression (Michael, 2018b, 2019a)[18]. Gulpek et al. (2014) found that "[t]he level of alexithymia in conversion disorder patients, without any other psychiatric disorder, is higher than that of the healthy controls" (p. 300). In an independent study, Demartini et al. (2014) found that "alexithymia was present in 34.5% of patients with (functional motor symptoms)" (p. 1132)[19]. This suggests that the inability to be conscious of emotions can lead to pathological symptoms, indeed the very kind of symptoms that first led Freud on the path towards psychoanalysis. Accordingly, Demartini et al. go on to propose that "one hypothesis is that some patients misattribute autonomic symptoms of anxiety, for example, tremor, paraesthesiae, paralysis, to that of a physical illness" (p. 1132). This is very much in line with my own Freud-inspired proposal about the causes of hysterical symptoms (Michael, 2018b). It suggests that, just as the trait inability to be conscious of emotions can lead to hysterical symptoms in alexithymics, so too it might be that the repression-induced inability to be conscious of certain emotions can lead to hysterical symptoms in non-alexithymics[20].

FREE-ENERGY AND THE PROCESS OF REPRESSION

While the above account of the repression of the consciousness of emotion provides an outline of the form that such repression can take, we have yet to describe the process of repression itself. Coming up with such an account presents some *prima facie* problems, the most pertinent of which is avoiding a "homunculus" interpretation. This is the problem of explaining a particular mental process, such as repression, without treating some part of the brain as agent-like, in the sense of possessing psychological states and engaging in choices and actions – in other words, as an agent within the agent. Boag (2012) articulates this problem in his discussion of an influential account of repression based on Sullivan's (1956) model of *selective inattention*, in which awareness involves intensive concentration on a target to the exclusion of other stimuli. Boag (2012) argues against such an account as follows (p. 195):

> A single mind cannot be both exclusively aware of the target and also filtering incoming stimuli. Furthermore, the perceived "relevance" (or "irrelevance") of stimuli is a judgement, which cannot preclude both awareness and evaluation of target material (though this need not be conscious itself). Consequently, selective inattention here requires that all incoming material be screened to determine whether it is or is not relevant.

This brings home the problem in providing a neuroscientific account of repression: what we require is an account of the process of repression that avoids treating it as the act of some inner agency, that is, some homunculus in the brain. It is here that the free-energy perspective can be of most assistance, as we shall see.

A second problem relates to the question of the purpose of repression. Why would the brain-mind repress the consciousness of an emotion? The consciousness of emotion is presumably an adaptive state, providing for a considerably more flexible response to one's emotion than one would have if the emotion were unconscious. For example, it may be essential to the adaptive emotion regulation strategy of reappraisal (Subic-Wrana et al., 2014). Moreover, as we have discussed, it is probable that the absence of the consciousness of emotion often leads to psychopathology, such as hysterical symptoms. As such, it is, on the face of it, puzzling that there should be such an apparently maladaptive process as the repression of the consciousness of emotion. Once more, the free-energy perspective could be of assistance in addressing this question.

The free-energy perspective is useful for understanding what Freud called the "quantitative" dimension of mental activity. For Freud, that there is a quantitative dimension to mental activity is a fundamental tenet of his metapsychology, and he sought to understand all of the mind's dynamics in terms of this factor (Freud, 1950 [1895]/1957). For example, he posits that "the use of the terms "unconscious affect" and "unconscious emotion" has reference to the vicissitudes undergone, in consequence of repression by the quantitative factor in the instinctual

[18]The relationship between alexithymia and psychosomatic disorders has been recognised for some time, for example by Nemiah (1977) and McDougall (1982).

[19]A more recent designation of hysterical symptoms is as "functional neurological symptoms," of which a prevalent kind are "functional motor symptoms."

[20]It may be that there is a yet closer relationship between alexithymia and repression. Taylor et al. (2016) have noted a parallel between alexithymia and emotional repression, particularly with respect to Freud's notion of primal repression. The association between alexithymia and repression is also in line other studies: the imaging work of Lieburg et al. (2012, p. 665) indicates that alexithymia is associated with "higher connectivity in right-sided prefrontal regions" of the brain – regions that may correspond with repressive processes (Depue et al., 2007; Kikuchi et al., 2010) – a finding supported by more recent work (Kim et al., 2020).

impulse" (Freud, 1915/1957, p. 178). Elsewhere he refers to this quantitative factor as "psychical energy" (e.g., Freud, 1900/1957, p. 568) or as the "sum of excitation" (e.g., Breuer and Freud, 1893-1895/1957, p. 86). Such quantitative expressions have fallen into relative disuse in psychoanalysis (Akhtar, 2013, p. 14), partly because of the difficulty in applying them, and partly because they have been subject to criticisms on the grounds of not having any obvious neurobiological underpinning (McCarley and Hobson, 1977). Recently, however, there has been a revival of interest in this aspect of psychodynamics due to the work of Karl Friston. According to Friston and his co-author Carhart-Harris, "the [Freudian] process of minimising 'the sums of excitation' is exactly the same as minimising the sum of squared prediction error or free-energy in Helmholtzian schemes" (Carhart-Harris and Friston, 2010, p. 1270). By this, they wish to equate the key idea of the Bayesian brain hypothesis, that the brain seeks to minimise prediction error (or, on Friston's account, free-energy, which represents a bound on prediction error) with Freud's fundamental "principle of constancy," that the mind seeks to keep the level of psychical energy at a low and constant level.

The Bayesian brain hypothesis asserts that the brain is in the process of constructing hierarchically-organised multilevel "generative" models of the causes of sensory input, refining these in light of the input through Bayesian processes. At each level of the hierarchy of such a model, prediction units issue in predictions about the input from the level immediately beneath it, with the lowest level issuing predictions about the sensory input. These predictions are then compared, in prediction error units, to the input, and the difference, the prediction error, is fed up the hierarchy – thus the prediction error becomes the input to the next level. The inherent aim is to reduce the level of the prediction error, which can be done either by revising a model over a series of iterations (the basis of perception), or through bringing about movement that would change the sensory input in line with predictions (the basis of action). The theory is Bayesian because the processes by which predictions are generated correspond to those of Bayesian inference, in which the probability of a hypothesis is updated in light of evidence according to a formula involving the probability of the hypothesis prior to the given evidence – the "prior" – and the probability of the evidence given this hypothesis.

An important feature of this process is the role played by precision-weighting. This has to do with the degree of precision afforded to the prediction error versus the model at each level of the hierarchy. If more precision is given to the prediction error, then the model will be revised to a greater extent; if more precision is given to the model, then the prediction error will have less impact on revision. In cases where the model has an abnormally high precision, prediction error has little effect, and the representations given by the model become entrenched. This is, on my Bayesian account of hysteria (2018b), what purportedly happens with hysterical symptoms: a representation at a middle level of a hierarchical generative model, to the effect that the patient has a particular symptom, becomes entrenched due to excessively high precision being afforded to it, thereby coming to generate the symptom. On this account, the abnormally high precision is a consequence of the need to keep the real cause (an unconscious emotion) of changes in interoceptive input repressed.

The lowering of a model's precision can also be highly consequential. An example of this is given by Prosser et al. (2018), in their free-energy model of psychopathy. In this model they postulate three levels of "belief," corresponding to an unconscious self-schema (the lowest level), automatic conscious thoughts (the middle level), and high-level prior beliefs (the highest level). Importantly, the prior beliefs modulate the precision of the other two levels. It is through this modulatory connection that the authors account for psychopathic traits. For example, they model the psychopathic trait *lacks remorse* by having the prior beliefs lower the precision of a self-schema relating to feelings of shame or worthlessness. This leads to a relative decoupling of automatic conscious thoughts from such feelings, resulting in thoughts and behaviour that reflect the trait of lacking remorse. This nicely illustrates the pathological effects that the attenuation of precision can have on an agent.

I postulate that a roughly similar model can help explain the repression of the consciousness of emotions. In what follows I present only a preliminary sketch of such a model, as the details of a full model would be complex, taking us beyond the scope of the present paper. As in Prosser et al.'s model, there are, in this simplified model, three prominent levels at play. One, the lowest, corresponds to the experience of affect. The second, the middle level, corresponds to the second-order construal that constitutes the consciousness of the emotion. The third, the upper level, is a level superordinate to that of consciousness which modulates the precision of the levels beneath, that is, regulates consciousness. Such a superordinate level would correspond to a part of the Freudian ego, as it is the ego which, according to Freud, controls access to consciousness (Freud, 1926/1957, p. 95).

There is an important additional component to the model that has to do with the relation between the lowest level, pertaining to the experience of affect, and the upper level. In order to motivate this I turn to Connolly's (2018) suggestion about how we can understand Freud's "signal" theory of the triggering of repression from a free-energy perspective. Writing about situations of conflict between competing emotions, he proposes the following:

> In essence, the updating of the generative model after the first experience of conflict means that the conflict state itself becomes reflected at a superordinate level of organisation through the altered precisions. The sensory stimuli which would previously have generated the conflict state of uncertainty now generates the defence state that privileges one response over another. An example of such a response might be an inhibitory response of the prefrontal cortex towards the limbic system, which now occurs without necessarily reexperiencing the initial conflict state, but is rather the result of a downward prediction encoded at a cortical level. In essence the conflict is now "predicted" and "resolved" through one stroke, through the precision weightings towards one pole of the conflict now avoiding the uncertainty of the conflict state (p. 12).

The important points remain applicable even in the absence of direct conflict between competing emotions. In place of conflict, we may substitute a traumatic experience – corresponding to large amounts of prediction error[21] – brought on, in part, by the consciousness of an emotion. We may further suppose that, in the initial experience of the trauma, one of the means by which the prediction error was eventually reduced was by lowering the precision of the second-order construal that constitutes the consciousness of the emotion. If so, any future occurrences of that emotion could now come to generate the defensive response of lowering the precision on the second-order construal. That is, stimuli, such as a particular quality of affective feeling, that would previously have contributed to the generation of the second-order construal as an attempt to explain the feeling, now triggers (through prediction error feedback) a learned policy within the superordinate level of organisation (the third level of our model) for decreasing the precision of priors related to the consciousness of the emotion. This policy can be thought of as the operation of simultaneously predicting the re-experiencing of the trauma (hence large amounts of prediction error) and pre-empting it, in accord with the free-energy principle of minimising prediction error. Such a proposal, or an alternative that mirrors its general form even while differing in detail, enables us to avoid falling into the trap of positing homunculus-like agency to the brain, as there is no question of agency here, but rather simply a mathematically-governed process.

We can now turn to the second problem presented at the beginning of this section, recasting it in light of our free-energy model as follows: Why would the consciousness of an emotion elicit large amounts of prediction error? For, as mentioned, we may suppose that the consciousness of emotion plays an important role in the regulation of emotion, hence, if anything, would serve to reduce prediction error rather than increase it. An answer to the question is that the consciousness of an emotion can elicit high degrees of prediction error when it would be such as to lead to overwhelming negative affect, that is, affect that goes beyond that with which the brain can cope (hence warranting the epithet "traumatic"). Affect reflects prediction error (Solms and Friston, 2018), so overwhelming negative affect reflects a dangerous amount of prediction error.

This leads to an immediate follow-up question: Why would the consciousness of an emotion elicit such overwhelming negative affect? There are many possible answers, but I will focus on two that bear on important features of the consciousness of emotion.

The first possibility relates to my Bayesian account of hysteria. In this, the emotion whose unconsciousness leads to unexplained affect is unconscious due to being intimately connected with a repressed traumatic memory. We may relate this to the point made in section Levels of Emotional Awareness about the importance of situational context to the consciousness of emotion: for the emotion to become conscious, the situation that elicited that emotion would need to be accurately represented. In the cases we are considering, however, such situations have to do with memories that have been repressed, hence from the free-energy perspective have priors with low precision. Due to this repression they cannot be accurately represented, hence obstructing the construction of the second-order construal that would constitute the consciousness of the emotion. Indeed, going further, any attempt to make the emotion conscious would threaten the unconsciousness of the memory it is intimately associated with, so the policy of reducing the precision of priors associated with this traumatic memory may be extended to a policy of reducing the precision of priors associated with the consciousness of the emotion.

We may suppose that were this memory to become conscious, it would generate a degree of negative affect that would overwhelm the agent. Why so? On Freud's theory, such memories are subject to repression on account not just of the emotion immediately generated by the memory, but also due to deeper negative emotions associated with it, ones that potentially reach down into highly aversive childhood experiences or infantile sexual fantasies. Thus, the consequences of such memories becoming conscious are an escalating series of negative effects, corresponding to escalating amounts of prediction error. In order to prevent such a consequence, a policy is formed that reduces the precision of any priors related to that memory and its accompanying emotion, thereby preventing any such mental phenomena from entering consciousness. Such reduction in priors might not be enough, however, to prevent all affective consequences: the initial emotion of the traumatic event could still be stimulated. But repression prevents the second-order construal that constitutes the consciousness of such emotion from being produced, thereby holding back or ameliorating the escalating series of negative effects that would re-traumatise the agent.

The second possibility for why the consciousness of an emotion would elicit overwhelming affect relates more directly to my account of the consciousness of emotion as a second-order construal. If this account is correct, then such consciousness involves seeing *oneself* in a certain way (as having a particular perspective on some object). It is, in other words, a *self*-construal. In so being, it makes the consciousness of an emotion liable to impact on one's self-image, potentially bringing this into discord with one's ego ideal[22]. The consequences of such could to be bring about excessively harsh superegoic judgements about the self, leading to potentially overwhelming negative emotions. It is in order to prevent such emotions that the higher-level policy to reduce lower-level precisions is triggered. This relates to Freud's structural model of the mind, in which repression is seen to result from a conflict between superego and id. "Superego" here relates to high-level responses to one's self-construal, and "id" relates to the initial instinctual generation of the emotion.

[21] As Hopkins (2016) observes, "complexity [equal to free-energy plus accuracy, a measure of the predictive success of a model] is conceptually linked with *emotional conflict* and *trauma*." He goes on to explain that "experiences are rightly regarded as traumatic when the emotional adjustments (complexity) required for integrating them into thought and action are greater than the brain can manage."

[22] McDougall (1982) describes the conditions under which an ego ideal that is pathological in relation to emotions may develop. For example, as one of her alexithymic patients expressed, "In our family it was forbidden to be sad, or angry, or in need of anything. I still get confused if I try to grasp what I am feeling" (p. 84).

Thus, we may update our understanding of the process of repression as follows. Normally, the consciousness of an emotion is adaptive, as it helps in the regulation of the emotion (hence the reduction of prediction error). As such, normally the upper level of the generative model does not have a significant modulatory effect on the precision of the second level (or, perhaps, it increases the precision at that level). However, if in the past the agent has experienced overwhelming negative affect as a result (in part) of becoming conscious of the emotion in question, they develop, as a learned response, an alteration in the connections between the upper and the second level such that the precision of the second level is lowered in response to that emotion. In other words, a particular quality of negative affect has the effect of inducing the third level to lower the precision of the second level. The lowered precision at this level results in the failure of the emotion to attain consciousness. This leaves in its wake unexplained affect, but that is the price to pay for preventing the occurrence of the overwhelming affect which would have swamped the agent had the consciousness of emotion been allowed to develop.

DISCUSSION

The purpose of this paper has been to explore unconscious emotion in light of Freud's seemingly paradoxical remarks, in which, on the one hand, he claimed that consciousness was essential to emotion, and on the other, he frequently invoked unconscious emotion. My answer to the apparent paradox is twofold, with both solutions emanating from a particular philosophical account of emotion, namely, Roberts' account of emotions as concern-based construals. First, I pointed out an ambiguity in the concept of construal (reflecting an ambiguity in the concept of emotion) that allows us to give two slightly different accounts of emotion. In one, the narrow version, an emotion is constituted by the way one experiences an object, where this experience is coloured by the affect generated in response to the object. On this account, consciousness is essential to emotion. In the other, the broad version, an emotion is constituted by the organism's response to an object, where this response can be described as an evaluation of that object. On this account, consciousness is not essential to emotion. This latter account thereby allows, at least conceptually, for the possibility of emotion devoid of conscious experience.

The second and more important solution to the paradox draws on Roberts' account of what it means to feel an emotion. This account says that to feel an emotion is to experience oneself as construing an object in a particular way. If we equate this with the consciousness of an emotion, then we see how one can have an emotion without being conscious of it. This holds even if we adopt the narrow account of emotion described above, whereby consciousness – in the form of affective feelings – is essential to emotion.

This second solution opens up the possibility of the repression of emotion in a sense that goes beyond those which Freud spoke about, such as the suppression of the emotion. This is the repression of the second-order construal that constitutes the consciousness of the emotion. The existence of this form of repression is supported by evidence from alexithymia, a condition in which one can have an emotion without being conscious of it. It, moreover, complements my Freudian version of the Bayesian account of hysteria, for it is precisely due to the repression of the consciousness of an emotion that hysterics are left with the unexplained affect – hence prediction error – that leads to the formation of symptoms.

I further explored how this form of repression can be understood from a free-energy perspective, and thus addressed objections related to homunculi and the adaptiveness of repression. On this account, repression is the result of an affective signal that triggers a learned higher-order policy for reducing the precision of priors associated with the consciousness of the emotion that produced that affective signal. The policy has been learned as a result of past experiences, in which the consciousness of that emotion generated overwhelming affect, hence large amounts of prediction error. This generation of overwhelming affect may be explained in numerous ways, though I have focused on two explanations which draw on important facets of the second-order construal that constitutes the consciousness of an emotion. First, interpreting the way one is experiencing or responding to an object (i.e., the first-order construal) requires an understanding of the situational context. In the case where the emotion is interwoven with a traumatic memory, this would entail accessing this memory in way that could re-trigger the layers of affect underlying the trauma. Second, a construal of how one is construing things is a construal of one's self, thus potentially bringing such a construal into discord with one's ego ideal. This discord could generate overwhelming affect, hence large amounts of prediction error, due to superegoic responses to such conflict.

The exploration undertaken in this paper was an attempt to integrate philosophical, psychoanalytic, and neuroscientific viewpoints in addressing a number of interesting problems. The solutions I have offered to these problems are tentative, inspired more by an intention to show how different perspectives can inform each other than by an intention to provide definitive answers, so naturally there is much more to be said about all these issues. I hope to have shown, at least, that such an integration could be a fruitful source of ideas for making sense of the complexities of the psychodynamic aspects of mental functioning.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

I would like to express my sincere thanks to Tim Fuller and the reviewers at *Frontiers in Psychology* for their helpful comments on drafts of this manuscript.

REFERENCES

Akhtar, S. (2013). "Introduction," in *On Freud's "The Unconscious"*, eds S. Akhtar and M. K. O'Neil (London: Karnac Books), 1–19.

Barrett, L. F. (2017). *How Emotions are Made: The Secret Life of the Brain*. London: Macmillan.

Boag, S. (2012). *Freudian Repression, the Unconscious, and the Dynamics of Inhibition*. London: Karnac Books.

Breuer, J., and Freud, S. (1893-95/1957). "Studies on hysteria," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. II, ed. J. Strachey (London: Vintage).

Carhart-Harris, R. L., and Friston, K. J. (2010). The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain* 133, 1265–1283. doi: 10.1093/brain/awq010

Connolly, P. (2018). Expected free energy formalizes conflict underlying defense in Freudian psychoanalysis. *Front. Psychol.* 9:1264. doi: 10.3389/fpsyg.2018.01264

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. London: Penguin.

Damasio, A. R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. London: Vintage.

Damasio, A. R., Damasio, H., and Tranel, D. (2013). Persistence of feelings of sentience after bilateral damage of the insula. *Cereb. Cortex* 23, 833–846. doi: 10.1093/cercor/bhs077

Demartini, B., Petrochilos, P., Ricciardi, L., Price, G., Edwards, M. J., and Joyce, E. (2014). The role of alexithymia in the development of functional motor symptoms (conversion disorder). *J. Neurol. Neurosurg. Psychiatry* 85, 1132–1137. doi: 10.1136/jnnp-2013-307203

Depue, B. E., Curran, T., and Banich, M. T. (2007). Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. *Science* 317, 215–219. doi: 10.1126/science.1139560

Edwards, M. J., Adams, R. A., Brown, H., Parees, I., and Friston, K. J. (2012). A Bayesian account of 'hysteria'. *Brain* 135, 3495–3512. doi: 10.1093/brain/aws129

Eickers, G., Loaiza, J. R., and Prinz, J. (2017). Embodiment, context-sensitivity, and discrete emotions: a response to Moors. *Psychol. Inq.* 28, 31–38. doi: 10.1080/1047840x.2017.1255492

Freud, S. (1900/1957). "The interpretation of dreams (second part)," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. V, ed. J. Strachey (London: Vintage).

Freud, S. (1905/1957). "Fragment of an analysis of a case of hysteria," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. VII, ed. J. Strachey (London: Vintage).

Freud, S. (1909/1957). "Notes upon a case of obsessional neurosis," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. X, ed. J. Strachey (London: Vintage).

Freud, S. (1910/1957). "The future prospects of psycho-analytic therapy," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. XI, ed. J. Strachey (London: Vintage).

Freud, S. (1911/1957). "Psycho-analytic notes on an autobiographical account of a case of paranoia," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. XII, ed. J. Strachey (London: Vintage).

Freud, S. (1915/1957). "The unconscious," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. XIV, ed. J. Strachey (London: Vintage).

Freud, S. (1919/1957). "The uncanny," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. XVII, ed. J. Strachey (London: Vintage).

Freud, S. (1926/1957). "Inhibitions, symptoms, and anxiety," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. XX, ed. J. Strachey (London: Vintage).

Freud, S. (1933/1957). "New introductory lectures," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. XXII, ed. J. Strachey (London: Vintage).

Freud, S. (1950 [1895]/1957). "Project for a scientific psychology," in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. I, ed. J. Strachey (London: Vintage).

Green, A. (2004). Thirdness and psychoanalytic concepts. *Psychoanal. Q.* 73, 99–135. doi: 10.1002/j.2167-4086.2004.tb00154.x

Gulpek, D., Kelemence Kaplan, F., Kesebir, S., and Bora, O. (2014). Alexithymia in patients with conversion disorder. *Nordic J. Psychiatry* 68, 300–305. doi: 10.3109/08039488.2013.814711

Herrera, M. (2010). Representante—representativo, représentant—représentation, ideational representative: which one is a Freudian concept? On the translation of Vorstellungsrepräsentanz in Spanish, French and English. *Int. J. Psychoanal.* 91, 785–809. doi: 10.1111/j.1745-8315.2010.00306.x

Hopkins, J. (2016). Free energy and virtual reality in neuroscience and psychoanalysis: a complexity theory of dreaming and mental disorder. *Front. Psychol.* 7:922. doi: 10.3389/fpsyg.2016.00922

James, W. (1884). What is an emotion? *Mind* 9, 188–205. doi: 10.1093/mind/os-IX.34.188

Kikuchi, H., Fujii, T., Abe, N., Suzuki, M., Takagi, M., Mugikura, S., et al. (2010). Memory repression: brain mechanisms underlying dissociative amnesia. *J. Cogn. Neurosci.* 22, 602–613. doi: 10.1162/jocn.2009.21212

Kim, N., Park, I., Lee, Y. J., Jeon, S., Kim, S., Lee, K. H., et al. (2020). Alexithymia and frontal–amygdala functional connectivity in North Korean refugees. *Psychol. Med.* 50, 334–341. doi: 10.1017/s0033291719000175

Lacewing, M. (2004). Emotion and cognition: recent developments and therapeutic practice. *Philosophy Psychiatry Psychology* 11, 175–186. doi: 10.1353/ppp.2004.0054

Lacewing, M. (2007). Do unconscious emotions involve unconscious feelings? *Philos. Psychol.* 20, 81–104. doi: 10.1080/09515080601023402

Lane, R. D., Quinlan, D. M., Schwartz, G. E., Walker, P. A., and Zeitlin, S. B. (1990). The levels of emotional awareness scale: a cognitive-developmental measure of emotion. *J. Pers. Assess.* 55, 124–134. doi: 10.1207/s15327752jpa5501%262_12

Lane, R. D., Weihs, K. L., Herring, A., Hishaw, A., and Smith, R. (2015). Affective agnosia: expansion of the alexithymia construct and a new opportunity to integrate and extend Freud's legacy. *Neurosci. Biobehav. Rev.* 55, 594–611. doi: 10.1016/j.neubiorev.2015.06.007

Lange, C. G. (1885 [1922]). *Om sindsbevægelser: et psyko-fysiologisk Studie. Translated as The Emotions (along with William James "What is an emotion?")*, transl. A. Haupt (Baltimore: Williams & Wilkins).

Lazarus, R. S. (1991). *Emotion and Adaptation*. New York, NY: Oxford University Press.

Liemburg, E. J., Swart, M., Bruggeman, R., Kortekaas, R., Knegtering, H., Æurèiæ-Blake, B., et al. (2012). Altered resting state connectivity of the default mode network in alexithymia. *Soc. Cogn. Affect. Neurosci.* 7, 660–666. doi: 10.1093/scan/nss048

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

McCarley, R. W., and Hobson, J. A. (1977). The neurobiological origins of psychoanalytic dream theory. *Am. J. Psychiatry* 134, 1211–1221. doi: 10.1176/ajp.134.11.1211

McDougall, J. (1982). Alexithymia: a psychoanalytic viewpoint. *Psychother. Psychosom.* 38, 81–90. doi: 10.1159/000287617

McDougall, J. (1989). *Theaters of the Body: A Psychoanalytic Approach to Psychosomatic Illness*. London: Free Association Books.

Michael, M. T. (2018a). From Wittgenstein to Taoism: Philosophical applications of the concept of construal. *Cheolhak* 136, 83–108. doi: 10.18694/kjp.2018.08.136.83

Michael, M. T. (2018b). On the scientific prospects for Freud's theory of hysteria. *Neuropsychoanalysis* 20, 87–98. doi: 10.1080/15294145.2018.1544851

Michael, M. T. (2019a). The case for the Freud–Breuer theory of hysteria: a response to Grünbaum's foundational objection to psychoanalysis. *Int. J. Psychoanal.* 100, 32–51. doi: 10.1080/00207578.2018.1489705

Michael, M. T. (2019b). Self-Insight. *Int. J. Psychoanal.* 100, 693–710.

Nemiah, J. C. (1977). Alexithymia: theoretical considerations. *Psychother. Psychosom.* 28, 199–206. doi: 10.1159/000287064

Neu, J. (1977). *Emotion, Thought and Therapy: A Study of Hume and Spinoza and the Relationship of Philosophical Theories of the Emotions to Psychological Theories of Therapy*. London: Routledge and Kegan Paul.

Nussbaum, M. (1990). *Love's Knowledge*. Oxford: Oxford University Press.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford: Oxford University Press.

Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of Emotion*. Oxford: Oxford University Press.

Prosser, A., Friston, K. J., Bakker, N., and Parr, T. (2018). A Bayesian account of psychopathy: a model of lacks remorse and self-aggrandizing. *Comput. Psychiatry* 2, 92–140. doi: 10.1162/cpsy_a_00016

Pulver, S. E. (1971). Can affects be unconscious? *Int. J. Psycho Anal.* 52, 347–354.

Roberts, R. C. (1988). What an emotion is: a sketch. *Philos. Rev.* 97, 183–209.

Roberts, R. C. (2003). *Emotions: An Essay in Aid of Moral Psychology*. Cambridge: Cambridge University Press.

Scarantino, A., and de Souza, R. (2018). "Emotion," in *The Stanford Encyclopedia of Philosophy (winter 2018 edition)*, ed. E. N. Zalta (Stanford, CA: The Stanford Encyclopedia of Philosophy).

Smith, R., and Lane, R. D. (2015). The neural basis of one's own conscious and unconscious emotional states. *Neurosci. Biobehav. Rev.* 57, 1–29. doi: 10.1016/j.neubiorev.2015.08.003

Solms, M. (2013). The conscious id. *Neuropsychoanalysis* 15, 5–19. doi: 10.1080/15294145.2013.10773711

Solms, M. (2019). The hard problem of consciousness and the free energy principle. *Front. Psychol.* 9:2714. doi: 10.3389/fpsyg.2018.02714

Solms, M., and Friston, K. (2018). How and why consciousness arises: some considerations from physics and physiology. *J. Conscious. Stud.* 25, 202–238.

Solomon, R. (1980). "Emotions and choice," in *Explaining Emotions*, ed. A. Rorty (Los Angeles, CA: University of California Press), 251–281.

Stevens, F. L. (2016). The anterior cingulate cortex in psychopathology and psychotherapy: effects on awareness and repression of affect. *Neuropsychoanalysis* 18, 53–68. doi: 10.1080/15294145.2016.1149777

Stocker, M., and Hegeman, E. (1992). *Valuing Emotions*. Cambridge: Cambridge University Press.

Subic-Wrana, C., Beutel, M. E., Brähler, E., Stöbel-Richter, Y., Knebel, A., Lane, R. D., et al. (2014). How is emotional awareness related to emotion regulation strategies and self-reported negative affect in the general population? *PLoS One* 9:e91846. doi: 10.1371/journal.pone.0091846

Sullivan, H. S. (1956). *Clinical Studies in Psychiatry*. New York, NY: W. W. Norton.

Taylor, G. J., and Bagby, R. M. (2013). Psychoanalysis and empirical research: the example of alexithymia. *J. Am. Psychoanal. Assoc.* 61, 99–133. doi: 10.1177/0003065112474066

Taylor, G. J., Bagby, R. M., and Parker, J. D. (2016). What's in the name 'alexithymia'? A commentary on "Affective agnosia: expansion of the alexithymia construct and a new opportunity to integrate and extend Freud's legacy." *Neurosci. Biobehav. Rev.* 68, 1006–1020. doi: 10.1016/j.neubiorev.2016.05.025

Wittgenstein, L. (1953). *Philosophical Investigations*. Transl. G. E. M. Anscombe. Oxford: Blackwell.

Conflict of Interest: 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership