This PDF includes a chapter from the following book:

# Defining Mental Disorder

## Jerome Wakefield and His Critics

© 2021 Massachusetts Institute of Technology

**License Terms:**

Made available under a Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International Public License
https://creativecommons.org/licenses/by-nc-nd/4.0/

**OA Funding Provided By:**

The open access edition of this book was made possible by generous funding from Arcadia—a charitable fund of Lisbet Rausing and Peter Baldwin.

The title-level DOI for this work is:

doi:10.7551/mitpress/9949.001.0001

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

# 19 Harmful Dysfunction and the Science of Salience: Adaptations and Adaptationism

**Philip Gerrans**

Jerome Wakefield has proposed a definition of psychiatric disorder as involving an "inability of some internal mechanism to perform its natural function, wherein a natural function is an effect that is part of the evolutionary explanation of the existence and structure of the mechanism" (Wakefield 1992, 384).

The idea is that in psychiatric disorder, a mechanism is not performing the function it was selected for: either because of mechanistic malfunction or perhaps because current environmental conditions are too different from the environment that exerted selective pressure, causing a previously adaptive mechanism to have harmful effects (as when the human preference for high-calorie foods, which evolved as a response to scarcity, leads to diabetes in an environment of abundance). Wakefield is right to say that if we want psychiatric classification to reflect causation, we need to ensure that evolutionary considerations are theoretically incorporated in ways that help reveal underlying neural and cognitive mechanisms. There are, however, different ways to do this.

For strong adaptationists, the surface syndrome that confronts the clinician results from a problem with a cognitive adaptation whose nature can be inferred from the deficit. This, for example, is the aspiration of strong modular theory of mind deficit explanations of autism. The idea here is that some core deficits in social cognition characteristic of autism can be explained in terms of a developmental deficit in a modular capacity for mindreading that evolved as an adaptation for negotiating the social world (Cosmides and Tooby 1994; Baron-Cohen 1996).

It will, however, rarely be the case that a psychiatric disorder can be explained in terms of the (mal)function of a single well-defined cognitive adaptation. Patterns at the surface level, of behavior, belief, or experience, rarely directly reflect the operations of a single domain-specific cognitive system. More often, such patterns result from a cascade (often a developmental cascade) of events that ramify through the cognitive system (Karmiloff-Smith 1994, 1998; Stevens and Price 2000; Gerrans 2002, 2007; Stone and Gerrans 2006; Gerrans and Stone 2008). For this reason, strong adaptationist versions of evolutionary psychiatry are unlikely to succeed in directing us to cognitive or neural mechanisms. They are too "top down" in their analysis of the problem.

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

This is obvious in the case, for example, of theories that postulate an adaptive problem for which schizophrenic delusions represent a solution (Stevens and Price 2000; Dubrovsky 2002), but the point generalizes.

Once we abandon the strong adaptationist approach in favor of a cognitive neuroscience informed by evolutionary theorizing, psychiatric classification requires substantial revision. Wakefield's approach, I suggest, ultimately leads to abandoning the current *Diagnostic and Statistical Manual of Mental Disorders* (*DSM* ) approach in favor of that that recommended by Dominic Murphy (2006)—namely, to allow classification and psychiatric practice to reflect the architecture of the mind disclosed by cognitive neuroscience. I make my argument via a case study of a, perhaps the, classic psychiatric phenomenon: delusion. Once we focus on mechanistic explanation, the primary role of evolutionary theory will be in the explanation of mechanistic functioning rather than the definition of syndromes.

## I. Mechanisms of Belief Fixation?

It is especially difficult to preserve a taxonomic role for evolutionary theory in the case of psychiatric disorders like delusion, whose classification involves the concept of belief. This is so even though the human cognitive phenotype does show entrenched patterns of belief fixation in specific domains, and some psychiatric disorders can be characterized in terms of typical abnormalities in those patterns (e.g., social cognition is a specific domain and autism is characterized by deficits of belief fixation in that domain). However, those patterns are produced by low-level neurocognitive mechanisms that produce an upward cascade of effects ultimately expressed as patterns of belief fixation. It makes no sense to see these neurocognitive mechanisms, which regulate things like the allocation of cognitive resources to salient information, the anticorrelation of hemispheric activity, and ocular saccades, as cognitive mechanisms selected for forming particular classes of beliefs, although they do have drastic consequences for the type and content of beliefs we are able to form. Rather, they are mechanisms that enable processing of information at specific levels of cognitive complexity, and we can make progress on determining their nature by tracing their evolutionary history at the correct level of cognitive resolution.

Depression illustrates why classifying psychiatric disorders in terms of characteristic patterns of belief abstracts too far from mechanisms. Of course, the *DSM* does not characterize depression in terms of patients' beliefs (although it does refer to ideas, thoughts, and feelings, all "surface-level" phenomena), but the relationship between beliefs and ultimate causes of depressive disorder exemplifies the point I want to make. The patterns of beliefs of severely depressive patients are typically self-accusing, introjective, and profoundly negative, and they express an experience and expectation of failure and hostility in the social world. This is not, however, because of the malfunction of

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

a system designed to produce positive beliefs (perhaps by introducing positive bias into a domain-specific social reasoning system) about personal functioning. Rather, the beliefs of depressive patients express the life experience of someone whose engagement with the world has been disrupted by changes to very low-level cognitive systems (Gerrans and Scherer 2013).

Catherine Harmer and colleagues have examined the relationship between cognitive processing and mood following the administration of selective serotonin reuptake inhibitors (SSRIs) to depressive patients. They found that after one week's administration of serotonin, which changes the balance of norepinephrinergic and serotoninergic activity in the amygdala, patients' amygdala response to masked fearful faces was reduced, and the responses of their facial fusiform areas to happy faces were increased. Patients' explicit judgments about emotional expressions also changed accordingly. Patients were more likely to correctly identify positive emotional expressions, for example. Memory for positive words also increased (for a discussion, see Harmer et al. 2009). These effects have now been demonstrated repeatedly (Harmer 2008; Harmer et al. 2009; Di Simplicio et al. 2012; McCabe et al. 2011). Crucially, these effects occur well below the threshold of explicit awareness. Mood, however, does not remit so quickly, suggesting to Harmer that "antidepressants are able to modify behavioral and neural responses to emotional information without any change in subjective mood. Moreover the changes in emotional processing can be seen across different stimuli types and extend outside conscious awareness" (Harmer et al. 2009, 105).

The point is that the ultimate mechanisms here (or at least those positively affected by treatment) are subcortical systems involved in things like the scanning of faces to extract emotionally salient information. These mechanisms no doubt were selected for, since for humans, the eye region especially transmits information vital to reproductive success. A human who cannot automatically detect and process information about conspecifics' intentions and attitudes is at a huge social disadvantage. Not only that, but when these mechanisms fail to perform their normal role, the psychology of the patient changes drastically. Typically, her experience changes and her beliefs about the world and herself also change to reflect that experience, leading to the profile characteristic of depression. But it is not correct to say that those changes result from changes in a mechanism selected to form beliefs about the patient's prospects in the world.

The more we learn about the deep causal and cognitive structure of many psychiatric disorders, the more we discover about problems with this type of processing. Schizophrenic patients and patients with severe personality disorder also exhibit abnormalities of gaze tracking and face scanning (Green et al. 2000; Green et al. 2003).

These types of mechanisms seem to be precisely the entities described by Wakefield. They are mechanisms selected to perform fundamental aspects of social cognition. However, they appear nowhere in the characterization of depression in the *DSM*, and it is unobvious how their role in producing symptoms could be incorporated as part of

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

the definition of the disorder without a complete reconceptualization of the nature of psychiatric classification.

## II. Delusion and the *DSM*

I now turn to the case of delusion, which I shall argue presents similar challenges once we focus on the causally relevant mechanisms. The explanation of delusion requires understanding the selective history of relevant mechanisms, but the properties of those mechanisms have little to do with the concept of belief used in its definition.

The *DSM-5* defines delusion as follows:

> Delusions are fixed beliefs that are not amenable to change in light of conflicting evidence.… The distinction between a delusion and a strongly held idea is sometimes difficult to make and depends in part on the degree of conviction with which the belief is held despite clear or reasonable contradictory evidence regarding its veracity. (American Psychiatric Association 2013)

This redefinition departs considerably from the definition in *DSM-IV*:

> A false belief based on incorrect inference about external reality that is firmly sustained despite what almost everyone else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary.… When a false belief involves a value judgment, it is regarded as a delusion only when the judgment is so extreme as to defy credibility. (American Psychiatric Association, 2000)

Philosophers, as well as clinicians, have pointed out problems with the *DSM* definitions. For philosophers and cognitive scientists, the issue is highly salient in debates about the rationality constraint on belief ascription and the relationship between that constraint and theories of cognitive architecture (Stein 1996; Bortolotti 2005, 2009). Clinicians remain acutely interested in a closely related question: whether the difference between delusional and nondelusional belief is a matter of degree or kind. On some views, delusions are toward the end of a continuum of beliefs, which range from accurate beliefs impeccably produced according to canons of procedural rationality to psychoses (Maher 1988, 1999). On other views, delusions are the "hallmark of madness" radically discontinuous with normal beliefs (David 2013). The question of continuity of course matters in the clinic because treatment will vary according to etiology. The problem is complex because delusions are heterogeneous phenomena, ranging from the circumscribed and monothematic neuropsychological delusions to delusions associated with mood disorders and the grandiose and paranoid fantasies of schizophrenia. Thus, the *DSM* term really names a syndrome whose symptoms need to be explained on a case-by-case basis.

This syndrome aspect is exacerbated by the fact that delusions are defined as beliefs. The concept of "belief" itself refers to a syndrome: a pattern of behavior and thought, which, as Ryle put it, "hang together on the same propositional hook" (1949, 135).

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

There is no reason a priori to think that two believers of the same proposition, delusional or not, would acquire their beliefs in the same way. Testimony, experience, explicit reasoning, and intuition based on tacit cognitive processes can all produce beliefs, and the same belief can issue in different behavior depending on context. Thus, delusional and nondelusional patients can both believe the same proposition; it is the way it is believed that renders a belief a delusion.

Thus, the aim of a theory of delusion is a precise characterization of the difference between delusional and nondelusional subjects in the *way they believe*. And both *DSM* definitions direct us to a crucial feature of the difference: conviction in the face of obvious counterevidence. The idea is that nondelusional people would revise that particular belief in the face of readily available evidence.

The *DSM-IV* and *DSM-5* definitions characterize this doxastic rigidity differently. The *DSM-IV* explicitly invokes the notion of "false belief" and "incorrect inference." The *DSM-IV* thus suggests that the delusional subject is making a reasoning mistake of some kind. The nature and extent of such a mistake, and the validity of importing normative epistemic notions into the characterization of delusion, have been subjects of controversy in cognitive neuropsychiatry and philosophy for three decades now. A recent example is the revived discussion of the applicability of Bayesian principles to the understanding of delusion.

The *DSM-5* potentially sidesteps the problems raised by explicit reliance on normative notions of rational belief fixation, even though it uses the concept of insensitivity to change in the light of "conflicting evidence," which does lend itself to Bayesian theorizing and the project of conceiving delusion as a (possibly degenerate) form of abductive inference (Coltheart et al. 2010). The reason is that there is no essential need to involve normative epistemic concepts in explaining the tenacity of delusional belief. Perhaps the tenacity can be explained in psychological, cognitive, or neurobiological terms without reference to failure of reasoning or hypothesis testing.

Faced with these problems, one can see why one might opt for an evolutionary explanation: perhaps delusional beliefs are produced by malfunction of a module or group of modules designed by evolution to allow us to form beliefs on specific topics? After all, delusions do seem to cluster thematically: erotomania, grandeur, reference, control, paranoia, and so on. Perhaps this domain specificity has an evolutionary explanation. Very likely it does, but inference from delusional content to cognitive architecture will not get it right.

This type of adaptationist approach inherits the goal of evolutionary psychology to render belief fixation cognitively tractable by conceiving of the mind as a set of domain-specific reasoning devices. The mind does not have to search across all hypotheses to explain evidence provided by perception but restricts the search to a small set of hypotheses wired in by evolution. Thus, for example, we are evolved to explain conspecific behavior in terms of (most likely) intentions directed at us. This makes sense. The

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

most relevant information for hominids is social, and false positives have small cost compared to mistakes. Paying too much attention to potentially relevant social signals such as gaze or posture is a benign misallocation of cognitive resources, but missing signs of aggression, alliance, or care can be disastrous.

The problem for adaptationism about belief rather than about domain-specific perceptual systems is that belief, as evidenced by what people say and do, is the output of the integrated functioning of a complex hierarchy of cognitive systems rather than the product of a single specialized system. The construction of abstract representations of reality that integrate sensory or perceptual information with background knowledge is a quintessentially domain-general ability.

In recent literature, we find concepts such as "reality testing" and "belief evaluation" proposed as descriptions of cognitive functions compromised in delusion (Gerrans 2013). These descriptions in effect restate fundamental principles of domain-general rational belief fixation as candidates for cognitive processes that have gone awry in delusion. As such, they are not incorrect: it is true to say that people with delusions are not testing their beliefs against reality or evaluating them for consistency with other beliefs. However, such analyses are really perspicuous redescriptions, rather than theoretical proposals, precisely because they cleave to the language of belief.

The same is true of another recent proposal, the "doxastic shear pin" hypothesis, which in effect proposes a mechanism for what Jaspers referred to as the fundamental reorganization of psychic life produced by delusion (McKay and Dennett 2009). Something that struck Jaspers and early asylum psychiatrists, which is missing from contemporary cognitive accounts, is the peculiar experience of the delusional state and the mesmeric effect it exerts. It seems that (some) delusions are compelled by experiences that, despite their implausibility, are so intense and absorbing that the subject reorganizes her mental life to fit the experience (Jaspers 1968). The idea behind the shear pin hypothesis is that perceptual or sensory processes generate experiences that are so overwhelming that higher-level processing simply cannot cope in the normal manner by trying to make them consistent with rest of the information available to the mind: "The delusion disables flexible, controlled conscious processing from continuing to monitor the mounting…error during delusional mood and thus deters cascading toxicity. At the same time, automatic habitual responses are preserved, possibly even enhanced" (Mishara and Corlett 2009, 531). The delusion is like a safety switch triggered by the mind to deal with a power surge of confusing and distressing information that threatens to short out higher-level processing.

The difficulty with all these proposals is not that they get the phenomena wrong but that they all rely on the language of belief, which is intrinsically agnostic about mechanisms.

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

## III. Mechanisms and Salience

In fact, however, quite a lot is known about mechanisms implicated in delusion, in the sense that some neural correlates have been identified. However, until the cognitive role played by those correlates is explained and linked to delusion, correlation cannot be transformed into causation.

For schizophrenic delusions, at least, dopaminergic activity has been strongly implicated. Dopamine synthesis during psychosis is abnormal; antipsychotic drugs such as haloperidol are dopamine agonists, and schizophrenic and neurotypical brains differ in the distribution and action of dopamine receptors as well as the activity of dopamine projections from the brainstem to cortical areas (Grace 1991; Murphy et al. 1996; Goldman-Rakic 1997; Gurden et al. 1999; Moore et al. 1999; Jay 2003; Seamans and Yang 2004; Howes and Kapur 2009). So the idea that dopaminergic activity has a causal role to play in the explanation of delusion is irresistible. But we need a theory that links the molecular and neural events to the inability to let go of beliefs triggered by highly anomalous experiences. The *DSM* definition does not help.

At this point, a Wakefieldian might appeal to the idea of an evolved system not performing the function it was selected for: but surely the dopaminergic system did not evolve under selective pressure to produce accurate beliefs. Dopamine regulatory systems exist in all chordates and precursors exist in nonchordates. In the mammalian central nervous system, the dopamine (DA) neurotransmitter systems are diversified. Yamamoto and Vernier note that

> DA acts to modulate early steps of sensory perception in the olfactory bulb and the retina, motor programming, learning, and memory, affective and motivational processes in the forebrain, control of body temperature, food intake, and several other hypothalamic functions as well as chemosensitivity in the area postrema and solitary tract, to cite only the main of the DA-controlled functions. Dysfunction of DA neurotransmission was initially shown in Parkinson's disease. … In addition, DA has now been shown to significantly contribute to the pathophysiology of several psychiatric disorders such as schizophrenia, addiction to drugs, or attention deficit with hyperactivity. (Yamamoto and Vernier 2011, 1)

This is just to point out that a simple adaptationist explanation of the role of DA in delusion, which implicates malfunction in a domain-specific belief fixation system, would be misleading. In fact, it would be a case of psychiatry trying to link neural correlates to symptoms via "the extensive (and expensive) [search] for … non-existent entities" (Halligan and Marshall 1996, 5).

Nonetheless, the dopaminergic system provides a paradigm case of the relevance of evolutionary theory to the explanation of cognition and psychiatric disorder. It also provides a poster child for the successful integration of formal learning theory, computational approaches to the mind, and neuroscience. Explanation of the role of the

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

dopaminergic system invokes two key concepts from computational theory as well as evolutionary ideas: predictive coding and salience (Egelman et al. 1998; Braver et al. 1999; Braver and Cohen 2000; Durstewitz and Seamans 2002).

Predictive coding theories treat the mind as a hierarchically organized cognitive system that uses representations of the world and its own states to control behavior. All levels of the cognitive hierarchy exploit the same principle: error correction (Gottfried et al. 2003; Clark 2013; Hohwy 2013a). Each cognitive system uses models of its domain to predict its future informational states, given actions performed by the organism or its subsystems. When those predictions are satisfied, the model is reinforced; when they are not, the model is revised or updated, and new predictions are generated to govern the process of error correction. Discrepancy between actual and predicted information state is called *surprisal* and represented in the form of an error signal. Error signals are referred to as higher-level supervisory systems. These systems generate an instruction whose execution will cancel the error and minimize surprisal (Friston 2003; Hohwy et al. 2008). The process iterates until error signals are canceled by suitable action.

Thus, on the predictive coding model, the role of cognition is to detect and correct prediction error. When the world and the model of the world being used by the organism (or subsystemic component) match, there is no need to take further action, cognitive or behavioral. This principle applies universally across species (even those with rudimentary control systems we might hesitate to call minds). Even unicellular organisms need to navigate toward nutrients and away from toxins, as well as to learn and remember optimal behavior.

Thus, the most important information for the mind is signals of surprisal or prediction error. Such information is the most *salient* for any cognitive system since it signals misalignment between what the organism is trying to do and what it is actually doing. Not only that, but once the error is corrected, the organism needs to remember that solution and update its model of the world accordingly.

This framework has proved essential to the interpretation of the functioning of the human DA system. The DA system is essentially a salience system: it signals which information is relevant and needs to be the focus of activity. It does so by selectively enhancing activity in neural circuits, which represent salient information, allowing that information to dominate control functions, until an adaptive response is produced (Kapur 2003).

A crucial aspect of DA function is that it solves a problem, formally demonstrable in learning theory and predictive coding models, which recurs urgently in the wild: the problem of *reward prediction*. Consider a foraging squirrel faced by two trees, an oak and a pine. Climbing is exhausting, and only oaks have acorns. Eating acorns is intrinsically rewarding; climbing is not. So the squirrel does not need a reward to learn to eat acorns, any more than humans need to learn to enjoy mother's milk or high-calorie food. As it explores its environment, it needs to learn which trees are worth climbing and install

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

the right, intrinsically unrewarding, instrumental behavior—namely, oak tree foraging. Were we to plant electrodes in the brain of a foraging squirrel, we might initially see activity in salience systems amplifying activity in sensory neural circuits activated by eating acorns. Over time, this activity would be replaced by activity in the salience system amplifying initiation of successful foraging. Ultimately, we would see a spike in the DA system when the squirrel saw an oak tree followed by a lesser spike when it found an acorn and no spike at all if no acorns were found after the climb. The role of a salience system is not to reward success but to *predict reward* for an organism (Schultz et al. 1997; Berridge and Robinson 1998; Gottfried et al. 2003; Heinz and Schlagenhauf 2010; McClure et al. 2003; Egelman et al. 1998; Smith et al. 2006).

In animals like rodents, this type of behavioral biasing is called *incentive salience* since it makes potentially rewarding object motivational magnets (Schultz et al. 1997; Berridge and Robinson 1998, 2003; Braver et al. 1999; Tobler et al. 2005; Kapur 2003; McClure et al. 2003; Smith et al. 2006). Human cognition uses the same dopaminergic salience system at all levels to target cognitive function by selectively enhancing activity in neural circuits processing potentially rewarding information (Braver et al. 1999; Braver and Cohen 2000; Abi-Dargham et al. 2002; Durstewitz and Seamans 2002; Egelman et al. 1998; Goldman-Rakic 1997; Grace 1991).

At the level of brute mechanism, DA enhances the signal-to-noise ratio (SNR) between communicating neural circuits. It does so via the interaction of at least two types of DA action. Phasic DA, delivered in short bursts, binds to D2 receptors on the postsynaptic membrane. It is rapidly removed by reuptake from the synaptic cleft and acts quickly. It is described as producing gating effects: determining which representations are allowed to interrupt and enter controlled processing. Gating is a spatial metaphor; "entry into controlled processing" refers to levels of activation sufficient to capture and retain attention, as well as to monopolize working memory and executive functions. A pattern of neural activation amplified and reinforced by phasic dopamine activity dominates other patterns of activation.

Tonic DA, which acts over longer time scales, accumulates in the synaptic cleft and binds to presynaptic DI autoreceptors triggering reuptake. This contrast between tonic and phasic activity is a ubiquitous neuroregulatory strategy. Tonic levels of a neurochemical are delivered and maintained at steady levels by slow, regular pulses of activity. Phasic activity is intense, staccato, and short-lived, interrupting the ongoing activity maintained by tonic levels.

Phasic and tonic DA are thus antagonists and have different effects on the circuits they afferent. Phasic DA, acting on prefrontal cortical (PFC) posterior circuits, produces a gating effect. It allows new activation patterns in the PFC-posterior circuitry to be formed, allowing representations of new stimuli. Tonic DA maintains an occurrent activation pattern, allowing a process to be sustained against interference or competition. Together, phasic and tonic dopamine provide a mechanism for the updating

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

and maintenance of representations in working memory and thereby bias higher-level information processing adaptively (Arnsten 1998). The hypothesis follows and is consistent with neural network models that the balance of tonic and phasic DA is responsible for the rate of turnover of representations in the PFC-posterior networks (Grace 1991) that represent information required for higher-level cognitive processes. "Tonic DA effects may increase the stability of maintained representations through an increase in the SNR of background versus evoked activity patterns. In contrast, phasic DA effects may serve as a gating signal indicating when new inputs should be encoded and maintained" (Braver et al. 1999, 317).

This reward prediction framework tells us that the balance of tonic and phasic dopamine delivery would modulate the salience of representations at different levels, influencing learning, memory, planning, and decision and motivation. Furthermore, since unpredicted activity, which constitutes surprisal, is most salient and likely to be referred to as controlled processing, phasic dopamine activity that interrupts ongoing activity should be associated with novelty.

These predictions are borne out in single neuron studies of the ventral tegmentum area (VTA) of rats in a variety of paradigms. For example, in a conditioning paradigm, in the learning phase, dopamine neurons fire for the reward (Waelti et al. 2001; Montague et al. 1996; Schultz et al. 1997). As the association is learned, firing for the reward is reduced, and dopamine neurons fire for the instrumental behavior. In other words, they predict reward (Waelti et al. 2001, 43). Firing of dopamine neurons is modulated by nonarrival of predicted reward "in a manner compatible with the coding of prediction errors" (Waelti et al. 2001, 43). These neurons also respond to novel attention-generating and motivational stimuli.

In other words, it seems that the role of the dopamine system is to focus cognition on relevant stimuli. Events consistent with predictions produce less phasic activity in the dopamine system than novel (i.e., unpredicted) and affectively valenced (good or bad for the organism) events. Not only that, but once the associations are learned, dopamine functions as a reward prediction system, increasing firing for instrumental activity but reducing firing if the reward does not arrive.

As new cortical systems were layered over older ones, enabling higher levels of control and more abstract forms of representation, they inherited the same problems of resource allocation and adopted preexisting mechanistic solutions. In fact, we can see higher cognition as cognitive foraging: a search through representational space for relevant information. Thinking about lying on the beach is time wasting in the office but a vital use of cognitive resources at the travel agency deciding between holidays in Tahiti and Phuket. Imagining dying of skin cancer is ridiculous time wasting in the office but sensible cognitive resource allocation at the beach.

Higher levels of cognition face the same problems of adaptive biasing and exploit the same ancient mechanisms (Gottfried et al. 2003) to recapitulate the temporal

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

structure of reward prediction. For example, neurons in the ventromedial prefrontal cortex, a structure implicated in almost all personal higher-level cognitive processes, are innervated by the dopamine system and exhibit the same patterns of interaction with it in learning tasks as lower-level systems.

## IV. Dopamine and Delusion

The idea that the mind is a hierarchical control system using predictive coding principles, which depend crucially on salience systems, has an important implication for the explanation of delusion. Those experiences that signal surprisal will naturally dominate high-level cognition since it evolved to enable adaptive responses to problems that exceed the processing capacities of lower-level systems. The salience system will interact with the neural circuits that refer these problems, amplifying their activity (increasing the "gain" is the technical term in neural network theory) and thereby making them the focus of attention.

Against this background, the explanation of the role of the dopamine system in delusion turns out to be not so much a discrete psychological puzzle but a piece in the larger puzzle of understanding the relationships between lower- and higher-level control systems and the salience systems that modulate them. A brief survey of recent work shows just how far a deep understanding of delusion in terms of the processing of salient information takes us away from its characterization in terms of beliefs.

Recent work on the dopamine hypothesis of schizophrenia has concentrated on the role of dopamine in the salience system, comparing levels of phasic dopamine delivery in conditioning and learning paradigms to that of normal subjects. The basic idea is that in psychosis, the "wrong" representations become salient, and relevant novel information is not processed. At low levels, this is reflected in attentional deficits; at higher levels, it is reflected in failure to allocate metacognitive function appropriately. At all levels, what counts as surprisal (prediction error) is different for the schizophrenic mind as a consequence of abnormalities in the way the salience system works. Summarizing a range of studies, Heinz and Schlagenhauf express a developing consensus: "The blunted difference between relevant and irrelevant stimuli and outcomes may reflect chaotic attribution of salience to otherwise irrelevant cues, an interpretation that is in accordance with the idea that chaotic or stress-induced dopamine firing can interfere with salience attribution in schizophrenia" (2010, 477).

The salience interpretation of the dopamine system theory provides a unifying explanation of features of schizophrenia, including the characteristic phenomenology of the prodromal period in which subjects feel that events or objects are extremely significant and/or that they are hypersensitive. As Heinz and Schlagenhauf (2010, 474) put it, dopamine dysfunction may be particularly prominent during the early stages of schizophrenia before delusional mood is transformed into fixed and rigid patterns of delusion.

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

Transient episodes of felt significance are not abnormal, but in delusional subjects, dopamine dysregulation ensures that their hypersalience gives representations of objects or scenes a halo of significance and ensures that they continue to dominate attention working memory and executive function (Di Forti et al. 2007; Moore et al. 1999; Abi-Dargham et al. 2002; Grace 1991; Howes and Kapur 2009; Braver et al. 1999; Broome et al. 2005).

Following Laruelle and Abi-Dargham (1999), Kapur describes dopamine as "the wind of psychotic fire" (Kapur 2003, 14), which ensures that activity in circuits referring and processing delusion-related information increases to levels that make reallocation of resources to nondelusional information impossible for the psychotic subject.

A delusion is a response to that constant referral of surprisal, a "top-down cognitive phenomenon that the individual imposes on these experiences of aberrant salience in order to make sense of them" (Kapur 2003, 15). Once adopted, the delusion "serves as a cognitive scheme for further thoughts and actions. It drives the patients to find further confirmatory evidence—in the glances of strangers, the headlines of newspapers, and the tiepins of newsreaders" (Kapur 2003, 16). The effect of hypersalience is to entrench the delusion.

Mishara and Corlett (2009) wed this idea to the doxastic shear pin hypothesis, suggesting that prediction error signals from sensory systems are generated, amplified, and referred up the control hierarchy in delusion in a way that simply floods the executive systems with hypersalient information they cannot deal with (Hohwy and Rajan 2012; Hohwy 2013b):

> Delusions…involve a "reorganization" of the patient's experience to maintain behavioral interaction with the environment despite the underlying disruption to perceptual binding processes.…The delusion disables flexible, controlled conscious processing from continuing to monitor the mounting distress of the wanton prediction error during delusional mood and thus deters cascading toxicity. At the same time, automatic habitual responses are preserved, possibly even enhanced. (Mishara and Corlett 2009, 531)

## Conclusion

Clearly, this is not the place to evaluate the neuroscience of schizophrenia, only to note that recent research converges on the idea that one important symptom (delusion) can be explained in terms of aberrant activity in the salience system, which amplifies and refers prediction errors that high-level control systems cannot cancel.

Wakefield is surely right that the explanation of this phenomenon involves the functioning of a salience system "designed" by evolution, which is either malfunctioning (making the wrong information salient) or, if it is functioning correctly (referring prediction error in the form of anomalous experience), warping the functioning of other systems with which it interacts. The explanation of salience outlined above makes use of evolutionary ideas all the way through.

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

Assuming that the science of salience is on track, how should the classification of psychiatric disorders involving the salience system proceed? It is worth noting, for example, that the salience system is implicated not only in delusion but also in addiction and attention-deficit/hyperactivity disorder. We could, in principle, reclassify these psychiatric disorders as members of a family of *disorders of the salience system*.

My own view is that Wakefield, by directing attention to mechanistic functioning and evolved cognitive architecture, may be advocating a far more radical approach to classification than he thought—one in which the everyday or folk conception of psychiatric phenomena is replaced entirely. Or perhaps, if not replaced, it will survive as something like the Mendelian concept of a gene: a bit of folk shorthand useful for introducing an entity in terms of its phenomenology but ultimately not part of scientific understanding. I think that the right approach here is the radical one: a complete reconceptualization of the phenomenon using the vocabulary of cognitive neuroscience, informed by, but not necessarily invoking, evolutionary theory.

## References

Abi-Dargham, A., O. Mawlawi, I. Lombardo, R. Gil, D. Martinez, Y. Huang, D. R. Hwang, J. Keilp, L. Kochan, R. Van Heertum, J. M. Gorman, and M. Laruelle. 2002. Prefrontal dopamine D1 receptors and working memory in schizophrenia. *Journal of Neuroscience* 22(9): 3708–3719.

American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed (text revision). American Psychiatric Association.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. American Psychiatric Association.

Arnsten, A. F. T. 1998. Catecholamine modulation of prefrontal cortical cognitive function. *Trends in Cognitive Sciences* 2(11): 436–447.

Baron-Cohen, S. 1996. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press.

Berridge, K. C., and T. E. Robinson. 1998. What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* 28(3): 309–369.

Berridge, K. C., and T. E. Robinson. 2003. Parsing reward. *Trends in Neurosciences* 26(9): 507–513.

Bortolotti, L. 2005. Delusions and the background of rationality. *Mind and Language* 20(2): 189–208.

Bortolotti, L. 2009. Delusion. In *The Stanford Encyclopedia of Philosophy,* E. N. Zalta (ed.), Winter 2013. http://plato.stanford.edu/archives/win2013/entries/delusion/. August 18, 2014.

Braver, T. S., D. M. Barch, and J. D. Cohen. 1999. Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry* 46(3): 312–328.

Braver, T. S., and J. D. Cohen. 2000. On the control of control: The role of dopamine in regulating prefrontal function and working memory. In *Control of Cognitive Processes: Attention and Performance XVIII,* S. Monsell and J. Driver (eds.), 713–737. MIT Press.

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

Broome, M. R., J. B. Woolley, P. Tabraham, L. C. Johns, E. Bramon, G. K. Murray, C. Pariante, P. K. McGuire, and R. M. Murray. 2005. What causes the onset of psychosis? *Schizophrenia Research* 79(1): 23–34.

Clark, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36(3): 181–253.

Coltheart, M., P. Menzies, and J. Sutton. 2010. Abductive inference and delusional belief. *Cognitive Neuropsychiatry* 15(1–3): 261–287.

Cosmides, L., and J. Tooby. 1994. Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition* 50(1): 41–77.

David, T. 2013. Delusions: Not on a continuum with normal beliefs. In *Imperfect Cognitions: Blog on Delusional Beliefs, Distorted Memories, Confabulatory Explanations, and Implicit Biases,* L. Bortolotti and E. Sullivan-Bissett (eds.). http://imperfectcognitions.blogspot.com.au/2013/10/delusions-are-hallmark-of-madness.html. August 20, 2014.

Di Forti, M., J. M. Lappin, and R. M. Murray. 2007. Risk factors for schizophrenia: All roads lead to dopamine. *European Neuropsychopharmacology* 17(suppl. 2): S101–S107.

Di Simplicio, M., R. Norbury, and C. J. Harmer. 2012. Short-term antidepressant administration reduces negative self-referential processing in the medial prefrontal cortex in subjects at risk for depression. *Molecular Psychiatry* 17(5): 503–510.

Dubrovsky, B. 2002. Evolutionary psychiatry: Adaptationist and nonadaptationist conceptualizations. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 26(1): 1–19.

Durstewitz, D., and J. K. Seamans. 2002. The computational role of dopamine D1 receptors in working memory. *Neural Networks* 15(4–6): 561–572.

Egelman, D. M., C. Person, and P. R. Montague. 1998. A computational role for dopamine delivery in human decision making. *Journal of Cognitive Neuroscience* 10(5): 623–630.

Friston, K. 2003. Learning and inference in the brain. *Neural Networks* 16(9): 1325–1352.

Gerrans, P. 2002. The theory of mind module in evolutionary psychology. *Biology and Philosophy* 17(3): 305–321.

Gerrans, P. 2007. Mechanisms of madness: Evolutionary psychiatry without evolutionary psychology. *Biology and Philosophy* 22(1): 35–56.

Gerrans, P. 2013. Delusional attitudes and default thinking. *Mind & Language* 28(1): 83–102.

Gerrans, P., and K. Scherer. 2013. Wired for despair: The neurochemistry of emotion and the phenomenology of depression. *Journal of Consciousness Studies* 20(7–8): 254–268.

Gerrans, P., and V. E. Stone. 2008. Generous or parsimonious cognitive architecture? Cognitive neuroscience and theory of mind. *British Journal for the Philosophy of Science* 59(2): 121–141.

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

Goldman-Rakic, P. S. 1997. The cortical dopamine system: Role in memory and cognition. *Advances in Pharmacology* 42: 707–711.

Gottfried, J. A., J. O'Doherty, and R. J. Dolan. 2003. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301(5636): 1104–1107.

Grace, A. A. 1991. Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: A hypothesis for the etiology of schizophrenia. *Neuroscience* 41(1): 1–24.

Green, M. J., L. M. Williams, and D. Davidson. 2003. Visual scanpaths to threat-related faces in deluded schizophrenia. *Psychiatry Research* 119(3): 271–285.

Green, M. J., L. M. Williams, and D. R. Hemsley. 2000. Cognitive theories of delusion formation: The contribution of visual scanpath research. *Cognitive Neuropsychiatry* 5(1): 63–74.

Gurden, H., J.-P. Tassin, and T. M. Jay. 1999. Integrity of the mesocortical dopaminergic system is necessary for complete expression of in vivo hippocampal–prefrontal cortex long-term potentiation. *Neuroscience* 94(4): 1019–1027.

Halligan, P. W., and J. C. Marshall (eds.). 1996. *Method in Madness: Case Studies in Cognitive Neuropsychiatry*. Psychology Press.

Harmer, C. J. 2008. Serotonin and emotional processing: Does it help explain antidepressant drug action? *Neuropharmacology* 55(6): 1023–1028.

Harmer, C. J., G. M. Goodwin, and P. J. Cowen. 2009. Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *British Journal of Psychiatry* 195(2): 102–108.

Heinz, A., and F. Schlagenhauf. 2010. Dopaminergic dysfunction in schizophrenia: Salience attribution revisited. *Schizophrenia Bulletin* 36(3): 472–485.

Hohwy, J. 2013a. *The Predictive Mind*. Oxford University Press.

Hohwy, J. 2013b. Delusions, illusions and inference under uncertainty. *Mind and Language* 28(1): 57–71.

Hohwy, J., and V. Rajan. 2012. Delusions as forensically disturbing perceptual inferences. *Neuroethics* 5(1): 5–11.

Hohwy, J., A. Roepstorff, and K. Friston. 2008. Predictive coding explains binocular rivalry: An epistemological review. *Cognition* 108(3): 687–701.

Howes, O. D., and S. Kapur. 2009. The dopamine hypothesis of schizophrenia: Version III—the final common pathway. *Schizophrenia Bulletin* 35(3): 549–562.

Jaspers, K. 1968. The phenomenological approach in psychopathology. *British Journal of Psychiatry* 114(516): 1313–1323.

Jay, T. M. 2003. Dopamine: A potential substrate for synaptic plasticity and memory mechanisms. *Progress in Neurobiology* 69(6): 375–390.

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

Kapur, S. 2003. Psychosis as a state of aberrant salience: A framework linking biology, phenomenology and pharmacology in schizophrenia. *American Journal of Psychiatry* 160(1): 13–23.

Karmiloff-Smith, A. 1994. Beyond modularity: A developmental perspective on cognitive science. *International Journal of Language & Communication Disorders* 29(1): 95–105.

Karmiloff-Smith, A. 1998. Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences* 2(10): 389–398.

Laruelle, M., and A. Abi-Dargham. 1999. Dopamine as the wind of the psychotic fire: New evidence from brain imagining studies. *Journal of Psychopharmacology* 13(4): 358–371.

Maher, B. A. 1988. Anomalous experiences and delusional thinking: The logic of explanation. In *Delusional Beliefs: Interdisciplinary Perspectives,* T. E. Oltmanns and B. A. Maher (eds.). Wiley.

Maher, B. A. 1999. Anomalous experience in everyday life: Its significance for psychopathology. *The Monist* 82(4): 547–570.

McCabe, C., Z. Mishor, N. Filippini, P. J. Cowen, M. J. Taylor, and C. Harmer. 2011. SSRI administration reduces resting state functional connectivity in dorso-medial prefrontal cortex. *Molecular Psychiatry* 16(6): 592–594.

McClure, S. M., N. D. Daw, and P. R. Montague. 2003. A computational substrate for incentive salience. *Trends in Neurosciences* 26(8): 423–428.

McKay, R. T., and D. C. Dennett. 2009. The evolution of misbelief. *Behavioral and Brain Sciences* 32(6): 493–510.

Mishara, A. L., and P. Corlett. 2009. Are delusions biologically adaptive? Salvaging the doxastic shear pin. *Behavioral and Brain Sciences* 32(6): 530–531.

Montague, P. R., P. Dayan, and T. J. Sejnowski. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience* 16(5): 1936–1947.

Moore, H., A. R. West, and A. A. Grace. 1999. The regulation of forebrain dopamine transmission: Relevance to the pathophysiology and psychopathology of schizophrenia. *Biological Psychiatry* 46(1): 40–55.

Murphy, B. L., A. F. Arnsten, P. S. Goldman-Rakic, and R. H. Roth. 1996. Increased dopamine turnover in the prefrontal cortex impairs spatial working memory performance in rats and monkeys. *Proceedings of the National Academy of Sciences* 93(3): 1325–1329.

Murphy, D. 2006. *Psychiatry in the Scientific Image.* MIT Press.

Ryle, G. 1949. *The Concept of Mind.* Hutchinson.

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

Seamans, J. K., and C. R. Yang. 2004. The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology* 74(1): 1–58.

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

Smith, A., M. Li, S. Becker, and S. Kapur. 2006. Dopamine, prediction error and associative learning: a model-based account. *Network: Computation in Neural Systems* 17(1): 61–84.

Stein, E. 1996. *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford University Press.

Stevens, A., and J. Price. 2000. *Evolutionary Psychiatry: A New Beginning*. 2nd ed. Routledge.

Stone, V. E., and P. Gerrans. 2006. What's domain-specific about theory of mind? *Social Neuroscience* 1(3–4): 309–319.

Tobler, P. N., C. D. Fiorillo, and W. Schultz. 2005. Adaptive coding of reward value by dopamine neurons. *Science* 307(5715): 1642–1645.

Waelti, P., A. Dickinson, and W. Schultz. 2001. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412(6842): 43–48.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47(3): 373–388.

Yamamoto, K., and P. Vernier. 2011. The evolution of dopamine systems in chordates. *Frontiers in Neuroanatomy* 5(21): 1–21.

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022

Downloaded from http://direct.mit.edu/books/book/chapter-pdf/1893887/9780262362931_c001800.pdf by guest on 08 August 2022