# What Will Psychiatry Become?

**Dominic Murphy**

**Abstract** Modern psychiatry aims at uncovering the causal structure of mental illness. I discuss two issues relating to this. First, the allure of reductionism, which goes along with a metaphysical commitment to levels of explanation that gets in the way of more promising approaches to psychiatric explanation. Second, I discuss the place of psychology within psychiatry, suggesting that we may need to develop new psychological concepts to do justice to neuroscientific developments, but that this might rob psychiatry of the ability to help patients understand themselves.

## Introduction

This paper is about the kind of science that psychiatry needs, and a plea to shake off a way of thinking that suggests one popular answer to that question. Many scholars seem to think that what psychiatry needs is genetics, and/or some sort of reductive neuroscience. While those approaches are definitely powerful, there is a lot they simply do not capture, and other thinkers have emphasized the need for psychiatry to employ many different scientific approaches altogether. I agree with the latter perspective, but I think that it can be misleading to put the point in terms of levels of explanation, as is often done. Levels of explanation are often either different ways of describing the same system, or else another way of talking about levels of constitution, with smaller units making up bigger ones within the hierarchy of nature that runs from atoms to organisms. A natural corollary of this hierarchical picture of levels is a reductionist agenda that concentrates on lower levels in the hierarchy. I will call sometimes call this a "levels-based" approach. For various reasons, I don't think this suits psychiatry. I will say why, and investigate some of the consequences.

D. Murphy (✉)
Department of Philosophy, University of Sydney, Sydney, Australia
e-mail: Dominic.murphy@sydney.edu.au

P. Huneman et al. (eds.), *Classification, Disease and Evidence*, History, Philosophy and Theory of the Life Sciences 7, DOI 10.1007/978-94-017-8887-8_3,

Whatever particular sciences we expect to find contributing to psychiatry, there is a consensus, among biologically inclined thinkers, about its trajectory; it will go on to more fully "discover the facts about how things go wrong with the psychology and biology of human beings" (McNally 2011, p. 216). I read this is as a commitment to a strong version of the medical model in psychiatry. It says that humans are made up of biological systems with a natural function, and science can discern these functions and say when they are not being discharged as nature intends. I have distinguished (Murphy 2013) this strong interpretation of the medical model from a weaker version. The weak interpretation is committed to gathering information about signs, symptoms, risk factors, treatments etc., but lacks the ontological commitments of the strong version, which sees mental disorders as pathologies of mechanisms of the nervous system. The weak version sees mental disorders as syndromes and is agnostic about their biological basis.

This essay will be concerned with the strong version of the medical model. It assumes that we can talk about some brains as dysfunctional and others as falling within biologically healthy ranges of functioning, and that the explanation of the abnormalities can be carried out using the vocabulary of some favoured sciences of the brain. After setting this out in a little more detail, I'll ask what those sciences might be. Despite some recent trends in psychiatry, I shall argue that nothing in the medical model requires us to restrict ourselves to purely biochemical explanations. Rather, I will argue that the logic of causal explanations in psychiatry make many types of explanation possible. This position makes reductionism less attractive, but it also raises questions about the sorts of cognitive theories that we might employ.

## The Medical Model

The strong version of the medical model (and from now on I'll just say "medical model") interprets mental illnesses the way biological disease has been seen since the nineteenth century; to wit, as departures from normal functioning in some biological system. So understanding disease means understanding the normal function of bodily systems, and in psychiatry that means the brain or the central nervous system. Mental disorders are realized in neurological systems that are not doing what they should. So we need to understand normal and pathological function in terms that make that failure perspicuous – we ask what the system has evolved to contribute to the overall system that it partly constitutes, and how it is failing to do that. As Thagard (2008, p. 340) puts it clearly:

> the circulatory system consists of a set of components—the heart, veins, arteries, and blood—that interact to provide nutrients to the rest of the body. This mechanism is susceptible to many kinds of breakdown, such as defects in the heart valves, blockage in the arteries due to plaque and blood clots, and abnormal growth of blood cells. These breakdowns can arise because of many kinds of interacting causal factors, from internal ones such as defective genes to external ones such as infectious agents.

> Similarly, the explanation of mental diseases requires specification of the normal functioning of the brain and other relevant organs, along with precise description of the different kinds of breakdown that can impede mental functioning.

This sets out the agenda very clearly, but obviously raises numerous questions. The very idea that science can tell us what goes wrong with people in such a straightforward fashion is itself very controversial, since many theorists contend that specifying the correct functioning of a biological system is not a simply scientific question. Every species exhibits variation – this idea is the essence of Darwinism, so we should expect all biological systems to vary across members of a species. The question, then, is how to give sense to the idea that there is some correct state in which a natural system should remain, in the absence of final causes or some other way of saying what nature ought to be like.

Kincaid (2008, p. 375) has argued that it is unreasonable to see the understanding of the normal function of biological systems as part of the medical model. We can investigate depression (his example) based on "partial and unsystematic" understanding of its causes, as we do with organs in medicine more generally. But Kincaid identifies the possession of background theory with having a "complete wiring diagram of the organism from fertilization to maturity" (p. 377). Furthermore, he sees the search for such wiring diagrams as reflecting a view of science as a search for laws of nature and natural kinds. But these commitments do not have to hang together so tightly: a background theory of what a system does can be quite vague, but without some understanding of a system's typical function it is difficult to see how we could reach the conclusion that there was something wrong with it.

Graham (2010, pp. 53–58) offers a different criticism of the medical model which turns on the (plainly correct) observation that most mental disorders are the product of several different causes rather than one exclusive one, or in his terms, a set of propensities rather than a "single main cause" (p. 55) as in bodily diseases like malaria. However, this objection can be met if we distinguish between the realization of a disease and its more proximate causes. Many diseases have a number of different possible causes that interact with genetic propensities; lung cancer is not just caused by smoking, for example, but also by inhaling various pollutants such as asbestos or coal dust. The different causes share the power to exert a destructive effect on the respiratory system via the replication of abnormal cells of various sorts. We lack a comparably detailed story for mental disorders, but the logic would be the same. The proponent of the strong medical model bets that the different causes of a mental disorder will tend to render a set of neurological systems abnormal in the same way across the affected population, even if the details of the cases vary according to accidents of biography. It is not in dispute that the subjective intentional life of the patient makes a difference to how mental illness is experienced and manifested in different people. The medical model's fundamental contention is that all these people, despite the varieties in their presentations, have something in common at the neurological level: their neuropsychological systems are disrupted in ways that we make sense of using the explanatory resources of the neurosciences, including cognitive or intentional concepts. The proponent of a levels-based approach further insists that all the causes that push the system into dysfunction

must be amenable to micro-reduction. That might seem plausible if we think only of the relation between brain systems and their components, but it will not work when we widen our focus to take in the other, more distal, causes.

A second complication that Thagard's passage raises is that of the relation between dysfunction and disease. It is generally agreed that even if science can tell us when a biological mechanism is not working properly, that alone does not justify calling someone diseased. The additional judgement requires a different sort of basis, and most scholars agree that it must come from the norms of the surrounding community. This judgement is easy to make and share in cases where extreme pain, suffering or risk of death are present. But in many other instances, including a lot of psychiatric cases, judgements are likely to be contested. For example, suppose you inform your doctor that you wish to have your right leg amputated below the knee. It's very common to assume that if you have that desire, you are basically crazy. But the desire shows up in people whose mental health is otherwise unimpeachable, and they sometimes suffer acutely from the presence of what they feel to be an extraneous and unsightly body part. Perhaps we should just treat them as we would treat someone who wants a face lift, breast alteration or substantial tattooing; as harbouring a desire for a dramatic bodily alteration that might be unusual, but is not evidence of mental disorder.

Judgements that somebody is sick, bodily or mentally, are a particular family of judgements of deviance. Since communal norms are so important to our existence humans are deviance-detecting animals, and we draw many distinctions among counternormative behaviour. Sometimes we call it criminal, sometimes it is seen as immoral or eccentric. It can even come to be admired, and if it is then it might shift norms in a new direction. Some norms are violated in a way that makes us see people as ill, and specifying exactly what those deviant phenomena are is a tricky business.

The interaction between these two problems – judgements of malfunction and judgements of deviance – lies at the conceptual core of philosophy of medicine. Many philosophers think that we can be objective about at least one of the two steps; science can tell us what has gone wrong with a person's biology, and then we can ask whether the effects of that dysfunction on a person's life are of the right sort, whether in nature or severity, to make a diagnosis of illness or disorder.

It may be that the right picture is the reverse of this. We make a judgement of illness and then look for a scientific legitimation of it by investigating the biology or psychology of the subject to find out what might be wrong. Thinking of the procedure in this second way suggests that it is our habit of policing deviance, not our attunement to dysfunction, that is driving the show. In either case, judgements of dysfunction are critical. The medical model, conjoined to the tradition that sees cognition as information-processing, places the cause of mental problems in the failure of neurological mechanisms to function as they should. I am going to set aside all the big questions about whether we can make judgements of natural function in the absence of final causes. We still face the question I want to explore in the rest of the essay. How can judgements of malfunction be made in a way that helps psychiatric explanation, classification and understanding?

## A Tradition of Computational Neuroscience

The specification of the system of interest and the ways it can break down need to mention whatever concepts are necessary for understanding. The components of the brain are systems that govern the cognitive, sensory and motor capacities of the organism. It is normal in the neurosciences these days to view these systems as processing information. Cognitive scientists now employ computational models based on conceptions of information processing developed in the middle of the twentieth century. However, the basic idea of the central nervous system as a computational system dates back to the late nineteenth century and the idea that biological relations among parts of the nervous system can be modelled mathematically as dynamic transformations of the weights assigned to energy levels in and between cells, so that the output of a neuron is a function of the inputs to it. Nervous energy flowing through the system was modelled as sensory information ultimately derived from the environment, and specific states of at least sensory systems could be correlated with external states of affairs. Associations between ideas were modelled as changes in the connections between brain cells. William James (1890/1981, pp. 616–617) for example, offers a toy model of memory as a graph, with vertices interpreted as interconnected "nerve-centres" and retention depending on the strength of the connecting edges, which James calls 'paths' located "in the finest recesses of the brains tissues". This conception of the nervous system informs Freud's early thinking about the mind (in his posthumously published *Project for A Scientific Psychology* (Freud 1895)) and was common among his teachers (Glymour 1992; Wollheim 1990, ch. 3). The nineteenth century nervous system was a computational system in which mental activity is a process that adjusts connections between cells and the energy levels within them. We should distinguish this wider tradition from more recent claims which are characteristic of cognitive psychology, viz. that thought is manipulation of symbols: physical entities with semantic and syntactic properties.

## Modern Computational Neuroscience and Psychiatry

The tradition I refer to predates modern theories of computation, but it can clearly be seen as a forerunner of connectionism. I have followed (Murphy 2006) a wing of the strongly medical psychiatric community in urging that psychiatry should adopt the methods of contemporary cognitive neuroscience, as they descend from the information processing tradition, in order to carry forward the research program contained in the ideas of the medical model, in which classification and causal explanation will be ultimately founded on the neurophysiological organization of the mind. This approach is only one way to apply the medical model, and makes a bet that cognitive neuroscience is able to account for psychological phenomena by treating them as computational processes (though not necessarily symbolic process, rather than

connectionist ones). Skeptics about computational approaches to cognition can adopt other neuroscientific applications of the medical model. The worry is that those rival approaches lack the resources to deal with cognitive processes.

It is people, not parts of their brains, that are psychotic. But the explanation of why somebody is psychotic will cite problems with neurological mechanisms like the executive system in dorsolateral prefrontal cortex and its relations – perhaps, a failure of inhibition – with cognitive systems that have evolved to subserve thought less tethered to reality. An explanation in terms of the physiology of cognition does not rule out a broader range of upstream factors as sources of the functional disruption.

Following Kraepelin, we can distinguish etiology and pathology. An explanation of why Jane undergoes a psychotic episode could make reference to her recent trauma, or a failure to negotiate certain developmental challenges and a reliance on very destructive defence mechanisms (such as massive splitting and projection). To fit in with the logic of the medical model in its strong guise, however, such processes would need to have, among their effects, a realisation of a destructive or dysfunctional disease process in the brain. The ensuing neuropathology is just what the disease amounts to, on the strong interpretation of the medical model. That does not mean that the pathology must always arise in the same way, but if mental disorder is brain disease then there must be in every case a neuropathology – an abnormal state of a neurological mechanism – that realises the disease.

These mechanisms are cognitive systems involved in the regulation of social behaviour – I will say more later about what "cognitive" is likely to mean in psychiatric contexts. The systems are parts of larger biological systems – ultimately parts of organisms or even societies – and this leads naturally to a reductionist approach. Because biological systems can be described in many ways, they seem to cry out for a treatment in terms of levels of explanation; the same entity can be given a cognitive, computational or molecular interpretation, and since these are interpretations of the same thing the reductionist impulse has a clear opportunity. Fundamentally, it can seem, all the higher levels are just expressions of the lower. Let me now move to the contemporary scene, and try to weaken the grip of this connection between neuropsychology and the metaphysics of reduction.

## Levels, Mechanisms and Reduction

An influential statement of this reductionist impulse comes in Oppenheim and Putnam's famous "Unity of Science as a Working Hypothesis" (1958). They argued that in principle, psychological laws could be reduced to statements about neurons, which could be reduced to claims about biochemistry which could be reduced to atomic physics, and thus we could have a successful "microreduction" of psychology to physics. A microreduction in their sense is the decomposition of the entities in the theory being reduced into the proper parts of the reducing theory. The hope, and bet, is that this reducing theory will be the theory of the very smallest bits of nature.

The Oppenheim-Putnam picture is a very powerful and natural portrayal of a vision of explanation tied to a vision of the world. There is of course a large and detailed philosophical literature on the ramifications of this pair of visions, but I will not go into all that here. I just want to draw attention to the grip that the overall picture, in whatever specific form, has exerted. The world is seen as a hierarchy of levels of entities, with small ones nested inside bigger ones, and ultimately explaining how things work involves showing how the higher levels emerge from the lower. As well as expressing the metaphysics that dominate modern science and philosophy, the picture also comports well with an idea of explanation as involving showing how things work – taking the bigger thing apart to reveal the workings within it- as well as suggesting how laws at one level can hold at other levels.

But figuring out what it takes to unify science is one thing, and explaining particular psychiatric phenomena is quite another. There is no reason to suppose that they must be explained reductively, if that means employing only the concepts of low level molecular neuroscience or genetics (Microreduction, which involves explaining in terms of the laws of basic physics, might be possible in principle but is still a fantasy). One ideal of explanation in science involves showing how things work in terms of their components. This ideal naturally fits psychiatry and other biomedical contexts, because there is a dearth of laws in those contexts. So the mechanistic picture can be seen as validating the Oppenheim-Putnam vision if we forego laws and think about levels of mechanisms. On this account, what a system does is explained by the operation of the smaller systems that it comprises. But the explanation can work without being embedded in the Oppenheim-Putnam picture of the world at all. We can just think of explanation as depending on showing the processes that cause a phenomenon of interest to happen. Those factors do not have to be "lower": they just have to show us what happened. The ideal of explanation is showing how some things make something else happen, but the things that do the explaining do not have to be parts of the explanandum, nor from lower levels of the natural hierarchy.

Psychiatric phenomena should be explained using whatever concepts are necessary to explain them, and nothing in the logic of the medical model rules that out. It is true that contemporary cognitive science assumes that the mind decomposes into components and shows how the components work in concert to produce behaviour. It is also undoubtedly true, though, that mental illnesses are complicated phenomena; they are mixtures of behavioural, psychological and physical signs and symptoms which appear to depend on many different causes. The same condition in different people also varies in length and severity. For example, your chance of suffering from major depression depends on many factors. Genes certainly make a difference, but so do factors like the extent of the child abuse you suffered, the state of your marriage and your history of substance abuse, as well as stressful environmental events like unemployment or bereavement (Kendler and Prescott 2006, p. 281). Reducing unemployment or bereavement, as Oppenheim and Putnam think of reduction, is a fantasy. But they can still play parts in a causal model.

The medical model need only talk about causes, without specifying what they must be in advance, let alone assuming that there will be a microreductive account

of causes. Genetics and long-term unemployment may come together to explain why somebody is depressed without providing a reductive picture in Oppenheim and Putnam's sense. Depressive episodes are not made up of genes and units of unemployment, even if they depend upon them; the hierarchical picture of nature is not a good fit here, and the mechanistic picture works only in a very extended sense. But we can still explain something in terms of interacting natural phenomena. I will come back to this in a moment, but I should note the bigger problem for the Oppenheim-Putnam picture in psychiatry.

## From Laws to Mechanisms

The real problem that psychiatry (and not it alone) poses for the hierarchical picture of levels is that we are not, in psychiatry, dealing with one phenomenon described in different ways. Underpinning Oppenheim and Putnam's account, and the conceptions of level-based explanations that descend from it, is the idea that the same natural process is in view at each level, so that we are always talking about the same process. The psychology is realised in the microphysics, so that really the psychology just is the microphysics, only at a very high level of abstraction. A similar assumption is built into Marr's (1982, pp. 24–5) distinction between three levels of explanation in cognitive science, which has had a huge impact in philosophy. Marr's highest level specifies the computational task accomplished by the system we are studying. This says what the system does, specified in terms of what it computes. The middle level describes the actual representations and algorithms that realize the computation. The lowest level tells us how brain tissue or other material substrate, such as the parts of a machine, can implement the algorithm. Building a machine to do what natural systems do is what Marr was really after.

Again, Marr's three levels are different representations of the same process, which in his research program was vision, understood as the construction of a 3D representation of the world from two-dimensional data. This picture tallies with the Oppenheim-Putnam one, and also Craver's (2007) picture of mechanistic explanation, which stresses causal relevance. Causal relevance in this sense tells you how it is that something happening at one level makes a difference at another level: it is because the lower-level system is a part of the higher-level system.

The worry is that this account will fail to do justice to the fact that in psychiatry distinct causal processes work on different levels but are also indicative of distinct phenomena that do not exhibit part-whole relations. Say you become depressed through the interaction of genetic load and sudden bereavement; the latter is not reducible to the former.

A mereological picture of levels does not have to be conjoined with a reductionist approach to explanation. At the same time as Oppenheim and Putnam, Kenneth Waltz had something very like a picture of levels of explanation in *Man, the State and War* (1959), only he called them "images". Waltz's problem was explaining why wars start, and he argued for three images each of which gave the machinery

for a different explanation. One is human behaviour – wars start because people are aggressive. A second is the nature of polities – wars start because of the internal dynamics of states. A third is the nature of the state system – wars start because of the threats and incentives faced by nations in a system of actors with no overall control. Waltz did not have the metaphysical preoccupations of a philosopher, but there are clear hints of part-whole relations among his images: people constitute states, which in their turn make up the state system. However, there is much less of a reductionist tendency in Waltz – the explanations at the level of the state and the state-system are regarded as autonomous, and some room is given for each of them, even though the metaphysics of states is part-whole.

Waltz is an exception to the harmony I noted above between the picture of the world as a metaphysical hierarchy and that of explanation as reduction. He, like Oppenheim and Putnam, wrote at a time when all explanation was taken to depend on laws – you explained a phenomenon by showing how it is only to be expected, given the laws. You find this in the clinical literature of the time too: Cronbach and Meehl (1955) assumed in their account of validity that a theory in psychiatry needed to be a network of laws. But that picture has come under attack in recent years with the rise of mechanistic accounts of explanation (Bechtel and Richardson 1993; Craver 2007; Schaffner 1993; Tabery 2004). I want to turn now to the question of how the mechanistic account fits with the picture of levels of explanation.

For example, suppose we want to understand the mechanism by which neurotransmitters are released (Craver 2007, pp. 4–6). This involves finding answers to questions such as: why does depolarization of an axon terminal lead to neurotransmitter release, and why are neurotransmitters released in quanta? The answers involve pointing out various entities, including various intracellular molecules, and showing how their properties allow them to act. The entities interact with each other to give rise to the phenomena that we want to explain. An explanation with these features is mechanistic. In recent years philosophers have stressed the way in which explanation in many sciences, above all the biological and cognitive, depends on finding mechanisms (Bechtel and Richardson 1993; Craver 2007; Schaffner 1993; Tabery 2009). Rather than seeing explanation as a search for laws, we seek the parts within a system of which the structure and activities explain the phenomena produced by the system. Philosophers disagree over exactly how to characterize mechanisms, but it is agreed that mechanisms comprise (i) component parts that (ii) interact to give rise to the phenomena of interest. It is generally agreed that a mechanistic explanation shows how the parts and their interactions give rise to the phenomenon we want to explain.

The mechanistic picture also fits with a hierarchical or mereological understanding of nature. Mechanisms come in levels too, on the face of it: there are mechanisms in the cell that contribute to the larger systems that the cell is part of. Humans decompose into subsystems – reproductive, respiratory, cognitive – that decompose into organs, and it is easy to see these as levels of mechanism.

Central to Craver's account of mechanistic explanation, for instance, is causal relevance between phenomena at different levels of explanation (Craver 2002). Causal relevance is defined in terms of manipulability and intervention. Events at

one level are causally relevant in so far as they make a difference at another level. Causal relevance depends on realization. Levels of explanation, on this account, are, as we have seen before, actually descriptions of the same processes at different levels of resolution. A delusion can be understood in personal terms as a psychotic episode in the life of an individual that depends on relations between different psychological processes in different brain systems. These in their turn are involve cells whose operations can be studied in terms of the systems that constitute them, and on down to the molecular level. On this account, explanation in neuroscience, as in biology more generally, involves describing mechanism(s) at each level in ways that make apparent the relationships between causally relevant variables at different levels (Woodward 2010). Showing the causal relations between levels lets us integrate models of phenomena drawn from different areas of neuroscience. Clinical data, imaging studies and other high-level psychological information ultimately need to be systematically related to models of low-level phenomena such as the effects of neurotransmitter activity.

But again, we need to be careful in thinking of the biological hierarchy as licensing traditional levels-based thinking, because psychiatry does not deal with different interpretations of just one causal process, and its models look inherently multi-level in the sense that causal processes involve phenomena that span levels and are part of the same process, but are not just different ways of understanding one thing (Schaffner 1993, 2011 suggests that this is true throughout the life sciences). A complex causal structure at many levels is in a lot of ways a poor match for traditional level-based views, because the latter, I have suggested, has a natural tendency to reductionism.

Since the causes of many mental illnesses include a mix of genes and environmental factors we need to think about how environmental factors can be understood within the mechanistic program. These are different kinds of process, not different levels at which one process can be represented. If we were dealing with one process describable in different ways, then we could anticipate an integrative account in which higher-level variables get mapped on to lower-level ones. But even though it is hard enough to imagine a molecular or neurological reduction of a psychological construct it is even harder to imagine a reductive analysis of socio-cultural factors like unemployment or childhood sexual abuse. They have brain effects, but the brain effects vary across classes of individuals in ways that depend on other environmental and genetic contexts (see Kendler and Prescott 2006 for a comprehensive review.)

Appealing to levels of explanation is unobjectionable if it just involves a reminder that we need to relate variables of many sorts. But it is not clear that we have any principled grounds for sorting phenomena into levels, especially once we move beyond the organism: are unemployment and bereavement processes at different levels?

Marr did have a principled basis for distinguishing levels. He imagined them as descriptions of the same process (the construction of a 3D image from 2D retinal impacts) couched in the vocabularies of different sciences. But when we move outside the skull and begin introducing environmental factors and other kinds of cause, the Marrian picture looks less plausible. The topmost level in the

Putnam-Oppenheim picture, for example, is the social. But there is no one way to represent "the social" and even on their terms it is hard to see how it could be one level. In Waltz, for example, the state is at one level (or 'image') and the state system at another. This works because we have a straightforward part-whole relation between states and the international system, but consider religion. Is it a part of the social level, together with (say) family life and the economy, or are these all different? And if they are different, is the difference one of levels of explanation, or something else? Religious, economic and other social phenomena all have an effect on your health, so how do we represent them in a levels-type view, whether based on laws or on mechanisms?

I have tried to identify two problems for a view that allies level-based thinking with reductionistic thinking. One is the fact that the reductionistic approach is a poor fit with higher-level phenomena that are autonomous, rather than just an abstract description of the micro level. Another is that in psychiatry (though not just there) we want to deal with social phenomena, and lack any clear criteria for applying levels talk in that context.

## Two Responses

One way to deal with these problems is to assume that we can ignore the outside world because information about it is represented in the brain. Then, if we can understand the process of information transmission in the brain, we can reduce that to the micro-level. Adolphs (2010) for example, assumes that social neuroscience begins with the transduction of social information, so that social factors are relevant only in so far as the system represents them. The mechanistic-reductionist program can go ahead. Methodological solipsism of this type will work as an explanatory strategy if we want to preserve the mechanistic understanding of the biological hierarchy in an enduring system. It will uncover proximate mechanisms and the causal relevance relations between them. It will not help us to isolate the relevant environmental factors and understand their effect on the organism (because we have to know what they are first to make the solipsistic strategy work).

A different option preserves much of the mechanistic approach but takes a different view of causal relevance and a more relaxed view of levels. Campbell (2008) argues for an interventionist approach to causation. This is the view that when we say X is a cause of Y we are saying that intervening on X is a way of intervening on Y (Woodward and Hitchcock 2003; Woodward 2003; Pearl 2000): manipulating one variable makes a difference to another. This is not a definition or analysis of causation in other terms, since it makes use of causal ideas – it just states that questions about whether X causes Y are questions about what would happen to Y if we did something to change X. Kendler and Campbell (2009) have argued that an interventionist model provides a rigorous way of articulating the idea that any combination of variables might characterize the causes of a disorder, whilst at the same time providing a clear test of what variables are actually involved, thus avoiding a

simple-minded holism that just says that lots of things are relevant. Kendler and Campbell advance a picture of psychiatric explanation that looks for control variables that make a difference to behaviour, such as humiliation or genetic factors. But they do not expect to fit all the variables into a natural hierarchy in which events at one level are reductions of events at a higher level. Indeed, their picture, like that of the theory of causation it draws on, is silent about metaphysics. The point is that some set of variables can serve as what Campbell (2008, p. 209) calls the "control panel" for the system; there are some variables whose manipulation has a large effect on the outcomes. The moral of this tradition is that although correlation does not equal causation, patterns of correlations do. Or rather, by manipulating some variables we can have systematic effects on the phenomena, and this justifies using causal language (as when, based on correlations, we say that smoking causes cancer) and offers us opportunities to intervene in the system.

On this picture, environmental processes are part of the overall explanatory system in their own right, not just qua representations in the hierarchy of brain systems. We can continue to look for causal stories at many levels of explanation in the brain, but we do not have to face the worry of reducing environmental factors. On Kendler and Campbell's story, unemployment is a genuine cause of depression in so far as it makes a systematic difference to depressed patients, even if there is no explaining unemployment in terms of a mechanism. It is a genuine cause of depression in virtue of its difference-making properties.

Those difference-making properties cross levels. Depression counts as a cause of something neurological in its own right, and not just in so far as it is mediated by mechanisms that realise it. That is, cause and effect are related across levels. Kendler and Campbell contend (2009, p. 997) that interventionism "permits the clear separation of causal effects from the mechanistic instantiations of those effects", thus directly confuting the approach favoured by Craver and Bechtel (2007, p. 554) who argue that it stretches the concept of causation to breaking point to admit interlevel causes: they say that "to accept interlevel relationships as causal violates many of the central ideas associated with the concept of causation". Craver and Bechtel argue that we explain effects in terms of interlocking parts, and the relation across levels, they affirm, is one of constitution, not causation; causation can only be intralevel. Events at a level cause subsequent phenomena at that level. They in their turn realise higher-level phenomena. Interlevel causation, on this view, amounts to something causing itself, because different levels are different ways of talking about the same thing. Craver and Bechtel take causal relevance to be the relation borne by phenomena at one level to the lower-level phenomena they depend on, but causal relevance is not causation (Note that this notion of causal relevance is not the idea that something is a partial cause of a phenomenon (for which, see Northcott 2012)). The dispute here turns in part, then, on philosophical views about the nature of causation. Campbell (2008) argues on broadly Humean grounds that that we simply cannot tell in advance of inquiry what causal relations obtain in nature. We simply have to take our causal relations where we find them, including interlevel ones. I think Campbell (p. 214) is correct to see a commitment among reductionists to a view of causation that requires physical contact among cause and effect. This is far

easier to imagine in biological processes than in the relations between psychology and unemployment.

Appealing to levels of explanation in psychiatry, then, can either be a reminder that we need to relate variables of many sorts in explaining the causes of disorder. Or it can represent a commitment to a seeing psychiatric explanation in terms of a biological hierarchy, with systems built up out of other systems. The debate over how to understand mechanisms and processes at different levels is partly an empirical one and partly bound up with philosophical views on reduction, explanation and causation.

The problem with Craver and Bechtel's view is that it is so natural to talk of cross-level causation in psychiatry. It really does seem to be the case that variables at different levels have an effect on each other. Their contention is that when this happens we are only entitled to talk about genuine causation when we have a mediation of the higher-level by a lower one. That is, it is not the state of being unemployed that interacts with your inherited depressive genes to make your mood pathological. Rather, what makes the difference is the physical basis of being unemployed.

There is obviously something correct about this metaphysically. On any naturalistic picture of the world, everything is ultimately physical, and so being unemployed is an ultimately physical phenomenon: being jobless must supervene on a complicated disjunction of states of the world that physics could describe. But it is just a fantasy to expect a molecular or microphysical theory of unemployment, and the restriction of proper causal language to intralevel relations looks unmotivated. To say that unemployment causes depression does not seem to violate any basic ideas about causation as far as I can see, even if it relates a social cause to a psychological effect. And as I noted above, in some cases we are simply not sure about how to identify levels: if I say that slavery caused the US Civil War, am I relating causes at one level or across two?

Let me say where I think the discussion has reached. I have argued that the medical model seeks causal explanations for mental illnesses, regarded as pathological states of neurobiological systems. This raises problems because of the diversity of the symptoms and causes that psychiatry seems to acknowledge. The question is how to explain and represent these pathologies and a natural way is to acknowledge that psychiatry involves multiple levels of explanation. However, the commitment to levels has emerged from a tradition that sees scientific explanation as a part of a bigger, reductive explanatory project. This picture may be metaphysically appealing but it does not fit psychiatry. It relies on the idea that we are dealing with a unique processes unfolding in a natural system which admits of description in several ways. But in psychiatry we are typically dealing with several processes that are describable in only one way. A part-whole reductionist materialism is fine as a philosophy of nature, but for the explanatory, epistemic projects of psychiatry it is of very little use. Instead, we should look for control variables, the manipulation of which have a robust effect on the system we are interested in.

So, my original question, what sort of science does psychiatry need, can be reframed as what family of control variables offer the most promise for intervention

and manipulation in psychiatric contexts. The reductionist answer is – molecular ones. But that answer now appears as an empirical one. It can't be defended as a general metaphysical commitment, because that is beside the point, and on the face of it, it is not correct. Many non-molecular variables look to be just as good candidates for sites of manipulation and intervention. But as I said, this is now an empirical question, to be decided by measuring intellectual progress rather than fidelity to a picture of the world.

My bet, for what it is worth, is that all sorts of variables, from all sorts of sciences, will turn out to be relevant. But this answer raises another issue which I want to explore in the rest of the essay. The search for control variables is a search for scientific concepts – ways of representing natural phenomena that fit into our scientific apparatus for controlling the world. In psychiatry, part of the final family of concepts is very likely to be cognitive, or more broadly psychological. This can seem humane- a reaffirmation of human experience, of the mind, in the face of a reductionist agenda that is often charged with alienating the mentally ill by treating them as mere machines. However, this optimism is only feasible if the psychology we develop remains tethered to ordinary categories of thought. There is good reason to think this may not happen, and that the psychology we end up with will be heavily revisionist. It may therefore be just as remote and alienating as a purely biological psychiatry. I will end up taking up this issue, because answering the question, what sort of science does psychiatry need, involves answering the question of what psychology will become.

## Psychology, Humanity and Science

The medical model tells us what our explanatory task is. We need to explain the observed causal-statistical network of signs and symptoms in a class of patients by identifying the mechanisms inside the organism and the external factors that affect them, either by developing a model of the natural hierarchy or (I have suggested) by developing a control panel of important variables. The reductionist impulse tells us that the explanatory theory we develop should draw on the resources of the very small – if not by employing microreductive concepts, then at least by employing molecular ones. But there is nothing in the logic of the medical model that requires explanations of any particular sort. The medical model enjoins us to search for the causes of mental illness, and those causes could be biological, cognitive, social, or anything else that enables us to explain and predict. The medical model is about establishing the right causal pathways, and this task can be done independently of any metaphysical commitment. An explanation does not need to be biological to be useful.

We can regard psychiatry, then, as a form of cognitive neuroscience. This formulation makes room for the existence of cognitive explanations; indeed, Bentall (2003) has argued that cognitive psychology will typically do more to explain psychotic symptoms than alternative approaches. However there are two complications

that we need to address before we can resolve that cognitive psychology plays a role in psychiatry. The first is whether the nature of cognitive neuroscience leaves room for psychology at all as an explanatory programme. The second is whether, if psychology does exist within cognitive neuroscience, it will look anything like the psychology we use in everyday life, and whether, if it does not, the result will be alienating rather than liberating.

The research program of cognitive neuroscience assumes that the privileged decomposition of the human mind will be physiological. It looks for brain systems that have cognitive jobs to do, rather than for abstract computational systems. The old cognitive science (exemplified by Marr's approach) assumed that one could disentangle human psychology into abstractly described computational processes whose comprehension required no knowledge at all of the underlying biology. This picture has changed as we have learned, at least in the human case, that the decomposition of psychological capacities marches in step with the identification of physical structures within the brain that realize the component capacities. Typically, crucial evidence is provided by the absence of cognitive abilities in a subject who has an anatomical or physiological deficit in some brain area. The process is one that, in Glymour's words (1992) discovers "cognitive parts"; the ensuing decomposition is a physiology that is intentionally described. We use psychological language to characterise what it is that connected regions of the brain do – what makes them a system with a function.

An influential recent treatment of addiction by Ross et al. (2008), for example, starts from a set of behaviours that trouble any view of humans as fundamentally rational agents. For one thing, addicts "reverse preferences". That is, they expend resources trying to stay clean, and they also expend resources on their addiction. Rational choice theories suggest two reasons why you might behave like this. One is an increase in the relative value of short-term rewards as you approach them, so that they get more attractive the closer they are in time. Second, you might simply underestimate the costs of withdrawal. Both of these seem to be true of humans, which raises some puzzles. First; if these properties are shared by humans, why aren't we all addicted to all our preferred activities? A second puzzle is more specific; if anyone should be able to reliably estimate the costs of withdrawal it's former addicts, because they have been through it already. Yet addicts are more likely to get addicted than the general population. Ross et al. think that a cognitive psychological model is just the wrong sort of thing to do the explanatory job here, because physiology solves the puzzles. Specifically, they appeal to the operation of the dopaminergic system and its interaction with other systems.

The dopaminergic system is of interest to Ross et al. because it: learns environmental cues that predict reward; estimates comparative values of rewards; directs attention to cues that predict reward; prepares the system to act on those cues. The system, as they present it, has a set of functions that are described in intentional terms. My point is that this is a description, not an explanation, of how the system operates. The explanation mixes chemical and environmental concepts. The ventral tegmental area and Pars compacta of the substantia nigra release dopamine in response to surprising magnitudes or learned contingencies. This implements

learning: a flood of dopamine (in nucleus accumbens) tells your reward system that whatever it was attending to was better than expected. This sets up a feedback loop to direct further attention and cue the motor cortex to take action. Ross et al. argue that these properties jointly predict a system that will be captured by unpredictable shifts in small magnitudes. What prevents widespread addiction even though we are all built like this is the existence of frontal and prefrontal circuits that inhibit impulsivity through the integration of cognition – which regulates input to the reward system and emotion – especially risk aversion.

The psychology in this theory describes the function of the systems. The real explanatory power comes from the nature of the dopamine system and its relationship to other systems. We can describe these systems in rough intentional terms- the (midbrain) learning system, the (frontal) executive system, but it is unclear whether we should regard these as explanatory at all, or just heuristics designed to convey the function of the physiological systems. The explanation in terms of physiology would seem to lose none of its force if the intentional language were removed, but would cease to exist if we stripped away the vocabulary of systems neuroscience.

So, although the logic of the medical model permits explanations using any concepts you like as long as they are useful, we have a case here in which the psychology seems to be playing second fiddle to the neuroscience. It may even be that the neuroscience will crowd out the psychology altogether, as some eliminitavists have suggested. Feyerabend (1963) raised questions about our capacity to reduce folk psychology to physiology, and argued that any successful materialist theory would undermine folk psychology, by showing that there was really nothing mental at all. On this picture, physiology will not reduce psychology so much as replace it, and we might wonder whether the sort of explanations anticipated by Ross et al. do the same, by showing us how an account that might have been put in psychological terms, to do with learning, habituation and impulse, can be reframed in purely neurological terms. Perhaps the sort of science that psychiatry needs will not include psychology at all, but not because psychology can be reduced to something else. Rather, psychology will just be shown to be explanatorily vacuous compared to the emerging neurosciences. I think this conclusion is premature, but I do think that the neurosciences will change psychology. They may do so in a way that has potentially serious implications for ordinary experience.

Some aspects of our psychology don't matter to us. If experts come and tell you that you don't know how your brain parses sentences or responds to pheromones, you might not be bothered. But other aspects matter a great deal – we all care about our memories, our emotional life or the sources of our behaviour, and we do not want to be told that we are systematically wrong about them, especially not if the truth is expressed in scientific language that is incomprehensible to us. The truth about fermentation might be hard to grasp, but it does not interfere with your drinking. The truth about love or belief might be more disquieting.

The more prominent eliminativists have questioned the scientific credentials of folk psychology on these grounds. Churchland (1981) did fret about the reducibility of folk psychology to neuroscience, but his scepticism centred on other criticisms. He argued that folk psychology could do nothing to explain many psychological

matters, including mental illness. He also charged folk psychology with stagnation, since it had not changed since classical antiquity, and that it was poorly integrated with other sciences. These properties – explanatory poverty, stagnation and parochialism – are grave defects in a scientific theory. Churchland took them to be evidence that folk psychology was overdue for replacement by a successor theory, which he assumed would come from neuroscience. Like ether or phlogiston, beliefs and desires should be discarded as the non-existent posits of an obsolete theory.

Stich (1983) also argued that cognitive science would do without folk psychological concepts, assuming it would be largely computational and employ syntactically individuated states rather than intentionally characterised representations, and so not concerned with intentional states at all. Our folk concepts, he ventured, are too vague to be scientifically useful. Stich's point here is that there are many cases in which it is unclear whether the concept of belief really applies at all. Tamar Gendler has recently (2008) advocated for supplementing our notion of belief with one of *alief*. An alief is an automatic state that has some belieflike features, exerting some control over behaviour and cognition, and typically in tension with belief.

Hume considers the case of a man hung from a high tower in an iron cage. He 'cannot forbear trembling', despite being 'perfectly secure from falling, by the solidity of the iron which supports him' (Hume 1978; 1.3.13, p. 150). Hume puts this in terms of general rules (see Serjeantson 2005) learned from experience, with one rule supplied by imagination – that great height is dangerous – set against another, drawn from judgement – that iron supports are secure. Rather than judgement and experience, Gendler puts things in terms of alief and belief, which highlights the tension –does the man in the cage really expect to fall? No. But he can't help thinking it, or at least imagining it.

I suspect that the real terrain is more complicated than the simple tensions in Hume's or Gendler's accounts; there are probably lots of distinct information processing types in the brain that have some of the stereotypical aspects of belief. But we may very well need to draw a distinction between "bottom-up" processes that exert unreflective control, and "top-down" processes that are more deliberative and effortful. The eliminativist tradition may have been right to put to the potential revolution that science might work on our culturally bequeathed psychology, but wrong to think of abolition rather than reform.

Let's think again about the addiction example. I said that the story that Ross et al. tell is one in which there is very little psychology, but that might be because the psychology we currently have lacks the concepts to fit what the science has discovered. Gideon Yaffe (forthcoming) tackles this issue, asking what the dopamine signal actually represents. What within our commonsense repertoire of folk psychological concepts fits the activity of the dopamine system? All the science tells us is that dopamine represents some X such that we want more rather than less of X. But what is X? Should we think of the phenomenon in question as one of liking, wanting or valuing, for example? Addicts, past research suggests, can want something without much pleasure out of it (i.e. without liking it). Yaffe's suggestion is that the correct interpretation of dopamine signals is that they represent value. Addicts want drugs and this gives them a reason to value drug-getting at the time of consumption. But,

as with the neuroscience of belief or the distinction between wanting and liking, we may find ourselves groping for new concepts as the neuroscience throws our traditional concepts into confusion. Scientific advances have often caused large-scale reforms of our culture's view of nature. The hard thing to accept is that it might have a similar effect on our view of ourselves. That would be an epistemic advance, but the worry, as I have said, is that these new vocabularies will deprive people of their ability to understand themselves by replacing a familiar vocabulary with a remote, scientific one. Psychiatrists will always need to be able to help people understand what they have become. The worry is that greater understanding of the mind will make it harder for us to explain people to themselves.

## References

Adolphs R (2010) Conceptual challenges and directions for social neuroscience. Neuron 65:752–67

Bechtel W, Richardson R (1993) Discovering complexity. Princeton University Press, Princeton

Bentall R (2003) Madness explained. Penguin, London

Campbell J (2008) Causation in psychiatry. In: Kendler K, Parnas J (eds) Philosophical issues in psychiatry. Johns Hopkins University Press, Baltimore, pp 196–216

Churchland PM (1981) Eliminative materialism and the propositional attitudes. J Philos 78:67–90

Craver CF (2002) Interlevel experiments and multilevel mechanisms in the neuroscience of memory. Philos Sci Suppl 69:S83–S97

Craver CF (2007) Explaining the brain. Oxford University Press, New York

Craver CF, Bechtel W (2007) Top-down causation without top-down causes. Biol Philos 22:547–563

Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52:281–302

Feyerabend P (1963) Mental events and the brain. J Philos 60:295–6

Freud S (1895) Project for a scientific psychology. In: Standard edition of the complete psychological works, vol I. Hogarth Press, 1953–74, London, pp 283–398

Gendler TS (2008) Alief and belief. J Philos 105:634–663

Glymour C (1992) Freud's androids. In: Neu J (ed) The Cambridge companion to Freud. Cambridge University Press, Cambridge, pp 44–85

Graham G (2010) The disordered mind. Routledge, New York

Hume D (1978) In: Nidditch PH (ed) A treatise of human nature. Oxford University Press, Oxford

James W (1890/1981) The principles of psychology. Harvard University Press, Cambridge, MA

Kendler K, Campbell J (2009) Interventionist causal models in psychiatry. Psychol Med 39:881–887

Kendler KS, Prescott CA (2006) Genes, environment, and psychopathology: understanding the causes of psychiatric and substance use disorders. The Guilford Press, New York

Kincaid H (2008) Do we need theory to study disease? Lessons from cancer research and their implications for mental illness. Perspect Biol Med 51:367–378

Marr D (1982) Vision. W.H. Freeman, San Francisco

McNally RJ (2011) What is mental illness? The Belknap Press, Cambridge, MA

Murphy D (2006) Psychiatry in the scientific image. MIT Press, Cambridge, MA

Murphy D (2013) The medical model and the philosophy of science. In: Gipps R, Fulford W (eds) Handbook of the philosophy of psychiatry. Oxford University Press, Oxford, pp 966–986

Northcott R (2012) Partial explanations in social science. In: Kincaid H (ed) Oxford handbook of philosophy of social science. Oxford University Press, Oxford, pp 130–153

Oppenheim P, Putnam H (1958) The unity of science as a working hypothesis. In: Feigl H et al (eds) Minnesota studies in the philosophy of science, vol 2. Minnesota University Press, Minneapolis

Pearl J (2000) Causality. Cambridge University Press, Cambridge, UK

Ross D, Sharp C, Vuchinich RE, Spurrett D (2008) Midbrain mutiny: the picoeconomics and neuroeconomics of disordered gambling. MIT Press, Cambridge, MA

Schaffner K (1993) Discovery and explanation in biology and medicine. University of Chicago Press, Chicago

Schaffner K (2011) A philosophical overview of validity. In: Kendler K, Parnas J (eds) Philosophical issues in psychiatry II: nosology. OUP, Oxford

Serjeantson R (2005) Hume's general rules and "The chief business of philosophers". In: Frasca-Spada M, Kail PJE (eds) Impressions of Hume. Oxford University Press, Oxford, pp 187–212

Stich S (1983) From folk psychology to cognitive science. MIT Press, Cambridge, MA

Tabery J (2004) Synthesizing activities and interactions in the concept of a mechanism. Philos Sci 71:1–15

Tabery J (2009) Difference mechanisms: explaining variation with mechanisms. Biol Philos 24:645–64

Thagard P (2008) Mental illness from the perspective of theoretical neuroscience. Perspect Biol Med 51:335–352

Waltz K (1959) Man, the state and war. Columbia University Press, New York

Wollheim R (1990) Freud. Cambridge University Press, Cambridge

Woodward J (2003) Making things happen. Oxford University Press, New York

Woodward J (2010) Causation in biology: stability, specificity, and the choice of levels of explanation. Biol Philos 25:287–318

Woodward J, Hitchcock C (2003) Explanatory generalizations, Part I: A counterfactual account. Nôus 37:1–24

Yaffe G (forthcoming) Are addicts akratic?: interpreting the neuroscience of reward. In: Levy N (ed) Addiction and self-control. Oxford University Press, Oxford