Collected essays and conference proceedings

Kelly Alexandra Roe

2006

Contents

Chapter 1

Some thoughts on pain

Presented to the Consciousness at the Beach Workshop. Hosted by the Australian National University.

It is common to distinguish between two aspects to the experience of pain. On the one hand there is something along the lines of a sensory aspect. On the other, there is something along the lines of an evaluative, affective, aversive, or motivational aspect. I want to focus on the experience of the second aspect from within the framework of a representational theory of consciousness. In particular I want to consider firstly, whether there is a distinctive phenomenology of this second aspect. Secondly, whether the phenomenology entails representational contents. And thirdly, the relationship between evaluative representational contents and affect, aversion, or motivation. I won't be concerned with the debate around whether the sensory aspect is representational. Instead, I'll just grant that it seems a plausible way to go. What I want to focus on is whether there are similar prospects for a representational account of the aversiveness of the experience of pain. I should also note that I will be concerned with a fairly weak version of representationalism. I take the weaker version of representationalism to be the thesis that phenomenal properties entail representational properties. I'll remain agnostic as to

whether in addition to this representation entails phenomenology. I'll start by attempting to motivate a distinction between their being two aspects of the experience of pain.

There seems to be a fairly intuitive sense in which pain is essentially an experience and an essentially aversive one at that. When someone is in pain there is typically something hurtful, unpleasant, or aversive about their experience. The characterisation of pain as a feeling that is awful or aversive seems to be part of our common sense conception of pain. There are, of course issues around whether pain must essentially be experienced as when we become distracted. If unconscious pains are possible then it seems that pain isn't essentially experienced at all. Still, there does seem to be something intuitively plausible about the experience of pain being essentially aversive in the sense that when we experience pain it is experienced as being intrinsically awful, bad, or aversive.

Dennett argues that there is support for this aversive aspect of pain in the neuro- biology of the brain. While his model is fairly complex I want to focus on just one part of it where he distinguishes between a 'high' and 'low' road for the processing of pain stimuli. He maintains that:

> One channel carries through the lower, phylogenetically older portion of the brain... and the other passes through the thalamus and is projected onto the... neocortex. The new high path... subserves fine-grained perception: location and characterisation of pain and other stimuli. The old low path is characterised by orthodoxy as the *aversive* system, the "motivational-affective processing" system. Orthodoxy is well buttressed by evidence in this instance, and this suggested separation of the hurtfulness or awfulness of pain from its other characteristics... 201-202'.

Dennett thus distinguishes between the neuro-biological processing of sensory representational features, and the 'motivational-affective processing' of the aversive features of pain. Dennett maintains that 'pains are abhorrent, at least usually'. He also notes, however, that subjects on morphine report

'that the pain continues (and continues to be *pain*), though they no longer mind it'. Subjects with lobotomy similarly report feeling intense pain but not minding it. Dennett seems to take this as indicating interruption in low road processing while high road processing continues fairly much as usual. These cases thus seem to support the high and low road as processing different aspects of the experience of pain. On Dennett's account it seems that the aversive aspect is inessential to pain in the sense that he accepts the subjects report 'that the pain continues' even though they don't experience it as aversive. While there is usually aversiveness the folk seem to be wrong about the aversive aspect being essential. Dennett also acknowledges, however, that there typically is an aversive aspect to the experience of pain.

Tye is interested in offering a representationalist account of pain and as such he is focused on the experience of pain and representational contents rather than the neuro- biological underpinnings for a distinction between two aspects to the normal experience of pain. Tye attempts to argue that the sensory aspect of pain is representational. The phenomenology entails that the person represents tissue damage of a certain quality (such as throbbing, or stabbing) at a location. He maintains that such sensory representational contents are non-conceptual and map-like and that they represent in a three dimensional array. This account corresponds most directly to Dennett's High road of processing.

Tye also considers that there is another aspect to the typical experience of pain where the phenomenology entails an evaluative representational content. He maintains that normally we not only represent the sensory aspects of quality and location of tissue damage, but we also represent that damage as bad. The evaluative aspect of pain corresponds to Dennett's low road. Tye follows Dennett in citing the cases of morphine and lobotomy to support (and indeed to motivate) his distinction between distinct sensory representational and evaluative representational aspects of the experience of pain. Tye maintains that these cases show that one can represent tissue damage without thereby experiencing the damage as aversive. Subjects say that they continue to feel pain even though they no longer find it aversive.

Tye and Dennett seem to agree that the sensory representational high road is essential to pain rather than the aversive or evaluative representational low road. They concur that people on morphine and people who have had a lobotomy continue to experience pain even though they do not experience its aversiveness. Segear disagrees with this conclusion, however, maintaining

> Pain and pleasure have affective components which are essential to the identities of these mental states... a pain in one's ankle carries a great deal of sensory information about one's ankle but also involves the 'intrinsic irksomeness'... or aversiveness which makes it an instance of pain. P 15

All three seem to agree that there are two distinct aspects to the normal experience of pain, however, and that these aspects can come apart when people have morphine or a lobotomy. It might be the case that there is a decision to be made about whether the sensory or affective aspect is more essential to the common sense conception of pain. While typically the affective aspect seems most salient or concerning to us the cases of morphine and lobotomy show us that people will track the sensory aspect when the affective aspect is absent Tye maintains that one cannot have the affective aspect in the absence of the sensory aspect, but this does seem to be controversial.

Tye maintains that the experience of the aversiveness of pain is a non-conceptual representation similarly to how he characterised the sensory aspect of the experience as non-conceptually representational. While one might judge the sensory aspect is bad for one this is different from the non-conceptual evaluative representation that Tye has in mind. He maintains that 'those whose pains are normal experience the same sort of disturbance, but now it is experienced by them as bad or unpleasant'. He claims that the 'badness' of pain is an objective property of the tissue damage in the sense that

> A bodily disturbance can feel bad without really being bad for one. Suppose, for example, that some diseased tissue is damaged with the result that the virus it is harbouring dies and the tissue is no longer apt to harm. [Normally tissue damage releases

> prostaglandins which] affect blood pressure in a negative way. The ensuing shift in the body landscape, occurring as pain is felt, is not good to the subject. It is a departure from functional equilibrium. And this the subject experiences'. In this way, pain is usually an emotional experience as well as a sensory one... (Tye In Defense of Representationalism)

It seems that there are two different ways that we could understand this passage. The first way would be to say that whether the tissue damage is bad or not is an objective property of the tissue damage. When one experiences the damage as bad and the damage is not harmful then this would be a case of misrepresentation. Another way, which might be more what Tye had in mind, would be to say that the affective aspect represents the badness of the prostaglandins being released. On this account if one experienced the affective aspect but prostaglandins weren't released then the affective aspect would be misrepresenting. I'm not sure whether prostaglandins are released in cases of phantom limb pain. If they are and the second reading is correct then the affective component of their experience wouldn't be misrepresenting.

In specifying the truth conditions for the representational contents of the experience of pain Tye maintains that

> Pains represent correctly IFF they are caused by bodily damage and cause an immediate reaction of dislike. But they can misrepresent on both counts and still remain pains. So it is not necessary for pains to occupy the standard causal role to count as pains.

I want to focus now on how there could be misrepresentation in cases where one experiences pain and yet fails to have an immediate reaction of dislike. It might be that when one experiences the sensory representational aspect of pain that state is meant to cause the affective representational aspect. Its failure to do this (as in the case of morphine) wouldn't seem to be misrepresentation so much as failure to represent the affective aspect, however. Another way we could go would be to say that the sensory representational

and affective representational aspects are independent from one another in the sense that one represents sensory properties while the other represents the prostaglandins and they function independently. It might be that if the affective representational aspect occurs but fails to cause an immediate reaction of dislike that misrepresentation has occurred. I'm a little unclear on how failure to cause a response of dislike counts as a case of misrepresentation, however.

Tye characterises the masochist (pp. 134-135)

> the felt quality of the *pain* is the same for both of us. I find the felt quality horrible and I react accordingly. He has a different reaction. Our reactions involve further feelings, however. I feel anxiety and concern. He does not. Here there is a phenomenal difference.

The sensory and evaluative aspects are thus considered to be the same in the normal case and in the case of the masochist. Tye maintains that any difference in phenomenology between the normal case and the case of the masochist is solely a difference in cognitive reactions to the same phenomenal experience of pain. Normally evaluative representation causes anxiety and concern with their associated phenomenology whereas the masochist lacks these usual cognitive reactions. This seems to imply that the relationship between the affective experience and finding the quality horrible is a contingent connection.

Another way that we may be able to describe the masochist would be to say that while the phenomenology associated with the sensory representational aspect is normal there is a phenomenal difference in the affective experience of pain and the masochist thus fails to evaluate the tissue damage or prostaglandins as bad. As such the masochist wouldn't be expected to find the pain horrible. This would suggest that the masochist would lack the evaluative aspect of pain comparably to the person on morphine, however. I'm not sure that this is plausible.

It might be the case that the masochist has the usual affective experience

and evaluative representational content but other considerations lead them to override expressing the usual aversive response. This would be one way of preserving a necessary connection between the phenomenology of pain and its intrinsic aversiveness, or the evaluative representational contents and their essential aversiveness.

There seems to be a sense in which typical experiences of pain are intrinsically aversive. While I allowed before that some experiences of pain lack this aspect, such as in cases of morphine and lobotomy, it seems that normally pain is experienced as intrinsically have this aversive aspect. If the role of the evaluative representation is to cause aversion and dislike then evaluating tissue damage as bad but not being adverse to it wouldn't seem to be a case of misrepresentation, so much as malfunction, however. Tye seems to think that you can represent evaluatively without aversion and thus the connection between them is contingent. It is possible to have an experience of pain that represents badness without being intrinsically horrible, or aversive.

Tye is happy to go with a functionalist account of emotions, however. He maintains:

> Part of what makes a given state an instance of anger is its effects on what the person wants and / or believes, and relatedly on how he or she behaves. Anger, for example, normally causes the desire or urge to act violently with respect to the perceived cause. Fear normally causes the impulse to flee. Any sensory state that did not play causal roles like these would not be classified as an instance of anger or fear' Tye 127-128

As such he seems to accept a functional account of emotions where it is essential to the identity of emotions that they play the specified functional role. He also allows that the phenomenology of emotion can come apart from the functional role, however. In this instance he claims

> one might conceivably feel just what normal people feel when they are anxious and yet not be anxious oneself, if, for example, one's state has no tendency at all to cause one to behave or react

> anxiously (due to very odd inner wiring). But arguably the feeling then would not be the feeling of *anxiety* p 130.

As such he seems to think that what the state causes is part of the identity of the state. This might be what he is getting at with respect to pain when he said that evaluating something as bad but not having the appropriate response constituted misrepresentation. Tye seems to go one way in the case of emotions making their identity dependent on their playing a functional role while another in the case of pains allowing that they can come apart from their functional role, however. I am unclear on why he goes one way with respect to the phenomenology of pain and the other with respect to the phenomenology of emotion. It seems to me that the link between pain and aversion would be more intimate than the link between anxiety and anxious behaviour.

It seems more plausible to think that pain would have arrived on the scene prior to anxiety and other emotional responses and I am interested in the prospects for explaining the experienced valence or the pleasurable and unpleasant aspects of our experience of emotions derivatively from the experience of pleasure and pain. Tye maintains that in the case of desire:

> I need not *feel* a desire for the desire to exist. Still we do often experience a feeling of being "pulled" or "tugged" when we strongly desire something'. Tye p. 4

This phenomenology of being pulled or tugged would seem to be intrinsically rather than contingently motivational. It might be that the affective phenomenology of pain is intrinsically irksome, horrible, unpleasant, motivational, or aversive where to experience that affective phenomenology is not only to represent tissue damage or prostaglandin release as bad or unpleasant it is also just to have an experience that is intrinsically irksome, horrible, unpleasant, motivational, or aversive. Within ethics some have held the thesis that representing something as wrong entails that one has motivation to prevent it to come to be. While I haven't explored the ethical literature on the subject this may be an interesting parallel to explore. I'm not too

sure whether the thesis is ultimately defensible but I have the intuition that the connection between the affective experience of pain and its horribleness or aversiveness is intrinsic. If the affective aspect to the experience of pain entails representational contents then I would think that either representing tissue damage or prostaglandin release as bad is intrinsically irksome or horrible or that there is a motivational aspect to the affective experience of pain that isn't adequately accounted for in representationalist theories of consciousness.

Chapter 2

Defining mental disorder

Presented to the Australian Association of Philosophy Conference. Hosted by the University of New England.

A fairly large literature has accumulated on the notion of 'disorder' and related notions like 'disability', 'illness', 'sickness', 'disease', 'distress' and so forth. Different theorists enumerate these notions and the relationship between them a bit differently. In the attempt to avoid a purely verbal dispute I'll follow what seems to be a recent convention of using the term 'disorder' though I will say a bit later to reassure people that I am getting to the heart of the debate.

Dominic Murphy states that 'mental disorder' is a term that is used in a variety of different contexts. In his book 'Psychiatry in the Scientific Image' he maintains that the scientific concerns can be carved off from the extra-scientific concerns. He then proceeds to focus on the scientific concerns. He identifies the scientific concerns as the project of finding out the nature and causes of mental disorder. The extra- scientific concerns include the legal notion of insanity, issues of moral responsibility, and therapeutic concerns such as that of involuntary treatment. While I'm not completely convinced

that the scientific concerns can be isolated from the extra- scientific concerns at the end of the day I will attempt to focus on the scientific concerns in this talk. This is related to my greater project of trying to offer a foundation for a science of mental disorder.

The official definition of mental disorder is provided by the American Psychiatric Association in the clinician's handbook the Diagnostic and Statistical Manual of Mental Disorders. The APA states that:

> ...each of the mental disorders is conceptualised as a clinically significant behavioural or psychological syndrome or pattern that occurs in an individual and that is associated with present distress (e.g., a painful symptom) or disability (i.e., an impairment in one or more important areas of functioning) or with a significantly increased risk of suffering death, pain, disability, or an important loss of freedom. In addition, this syndrome or pattern must not be merely an expectable and culturally sanctioned response to a particular event, for example, the death of a loved one. Whatever its original cause, it must currently be considered a manifestation of a behavioural, psychological, or biological dysfunction in the individual. Neither deviant behaviour (e.g., political, religious, or sexual) nor conflicts that are primarily between the individual and society are mental disorders unless the deviance or conflict is a symptom of a dysfunction in the individual, as described above.

The DSM view exemplifies a general position that Murphy has dubbed the 'two-stage view'. According to the two-stage view there are two kinds of facts that jointly determine whether an individual is mentally disordered. The first set of facts are characterised as non-evaluative or objective facts about malfunction or dysfunction. The second set of facts are characterised as evaluative facts about harms to persons. The DSM definition can be broken down into three parts and the first two seem to correspond to the two stages of the two-stage view. The first part of the definition talks about 'present distress', 'painful symptoms', 'disabilities', and 'sufferings'. The notion that seems to be relevant here is that of harm to persons. The second

part talks about 'behavioural, psychological, or biological dysfunction in the individual'. Facts about dysfunction are thought to be objective in the sense that they are not dependent on or determined by our values.

The rest of the definition, while not fully transparent, makes more sense when viewed in its historical context. Despite there being two editions of the Diagnostic and Statistical Manual prior to 1980 it wasn't until the third edition that there was an attempt to define mental disorder. Coopers discusses how the APA felt the need to define mental disorder in response to sustained pressure and critique from gay rights activists and the anti-psychiatry movement. The third part of the definition is an attempt to state that moral deviance is not sufficient for mental disorder unless the previous two conditions are met. It is basically intended as a direct denial of the claim made by some anti-psychiatrists that 'mental disorder' is nothing more than a label given to people who deviate from social or moral norms.

One thing that is notably lacking from the APA's account of disorder is an account of what makes a disorder psychiatric as opposed to neurological or 'mental' as opposed to 'non-mental'. All that is said about this in the introduction to the manual is that 'the term general medical condition is used merely as shorthand to refer to conditions and disorders that are listed outside the "mental and behavioural disorders" chapter of the ICD'. Since the International Classification of Diseases manual does not tell us the difference between general medical conditions and psychiatric disorders either I think that it is fair to say that this is merely shifting the problem.

The distinction between psychiatric and neurological disorders is very controversial. The main issue is whether disorders that are currently regarded as psychiatric will become subsumed under neurology with advances in neurobiology or whether there is something distinctively 'mental' about mental disorders. While this isn't an issue that I will focus on here it is important to note the assumption that psychiatry shares the notion of disorder with the rest of medicine. Mental disorders are thought to be kinds of medical disorders as is reflected in psychiatry's status as a specialist field within medicine. While some theorists have argued that mental disorder involves a different

sense of disorder or a metaphoric use of the term 'disorder' the assumption of the two-stage view is that it is the same notion of disorder that is in play. This is important because problems for the two-stage view aren't thought to be problems merely for psychiatry, they are thought to be problems for the rest of medicine as well. This is also important as counter-examples to the two-stage view make use of examples from both psychiatry and general medicine and in what follows I too shall make free use of examples from psychiatry and general medicine.

I now want to turn to Wakefield's 'Harmful Dysfunction' analysis of disorder. His account is another example of the two-stage view and while it differs a little from the DSM definition it is a clearer statement than that provided by the DSM. Another advantage to focusing on Wakefield's account is that he offers arguments for his view and he has received much attention for his sustained and somewhat artful defence of his definition. The Harmful Dysfunction analysis states, in a nutshell, that a disorder is a failure of the evolutionary function of a mechanism where that failure results in harm to persons. While he has written a number of papers in which he states and restates and subtly alters his arguments I think that the following is a charitable reconstruction of why it is that he thinks that we should accept his definition of disorder.

- P1) It is a-priori that disorder is due to an inner dysfunction that results in harm to persons. (At this stage he regards 'malfunction' to be a pre-theoretic, folk notion).
- P2) a. It is a-priori that it is a matter of scientific discovery what causal process fixes the functions and dysfunctions. b. It is a-priori that the relevant process is the causal-historical process that is the explanation for the mechanisms existence and maintenance in current populations.
- P3) It is a-posteriori that the relevant causal-historical process for fixing which effects are functions and malfunctions (for explaining the existence of the relevant mechanisms) is evolution by natural selection.
- C) Disorders are thus failures of an inner mechanisms evolutionary

function that results in harm to persons.

Wakefield's general project is part of a strategy that Fullford dubs 'the naturalisation cascade'. The notion is that while a term such as 'disorder' contains both evaluative and non-evaluative components the non-evaluative component can be reduced to another term such as 'dysfunction'. Ultimately the naturalisation project is thought to bottom out in purely causal processes. While there are a variety of terms on offer (for example illness, disability, disease, etc) and while different theorists locate them at different points in the naturalisation cascade this seems to be a general approach to attempting to provide an account of the scientific foundation of medicine in general and psychiatry in particular. The notion is basically that psychiatry can be grounded in medicine and medicine can be grounded in biology. The scientific foundation of psychiatry is thus ensured. While Wakefield doesn't say anything about the naturalisation cascade that Fullford outlined he does draw a parallel between his analysis of disorder and the causal-historical analysis of terms like 'gold' and 'water'. It is important to note that instead of naturalising 'disorder' directly, however, Wakefield attempts to naturalise disorder by way of naturalising function and dysfunction.

While many theorists have objected to Wakefield's Harmful Dysfunction account the majority of the objections have focused on dysfunction rather than on harm. While I shall also focus my critique on dysfunction before I do this I do want to say a few words about the relevant notion of harm. The first thing to note is that the motivation for the harm component comes from the idea that it seems plausible that a person could have a malfunctioning mechanism and yet not be harmed. An example of this would be someone who had a malfunctioning mechanism that resulted in a phobia of flying. If the person didn't need or desire to travel then the person wouldn't be harmed by their malfunction, however, and thus while they do have a malfunction they do not have a disorder. The main motivation for the inclusion of the harm component in the DSM was that irrespective of whether homosexuality was the result of malfunction it was not a mental disorder insofar as it did not result in harm. One issue that this example seems to raise, however,

is whether harms to the individual that result from the prejudice of society count as harms. It seems clear that much more needs to be said about the relevant notion of harm.

Another thing to note about the harm component is that whether an individual is harmed or not is thought to be an evaluative or normative matter. This might seem surprising and yet I haven't found any arguments as to why the relevant notion of harm is thought to be evaluative or normative. While it is often noted that whether a person is harmed by a malfunction can be highly dependent on their social and cultural environment this wouldn't seem to rule out the possibility that there are non- normative or non-evaluative facts about whether or not an individual in a certain socio-cultural environment is harmed and facts about how much they are harmed. There does seem to be a lot of work to be done on the notion of harm before we have a satisfactory account of disorder. I won't tackle this issue here, however. What is important to note for my purposes today is that the two-stage view maintains that there is a normative or evaluative component to disorder but that that is completely separate from the non-evaluative notion of dysfunction. I now wish to turn to a critique of the dysfunction condition and I'll restrict my criticism to the dysfunction component of the two-stage view.

It was once thought that xx (or female) was a malfunctioning sex in the sense that females were malfunctioning males. This was a particularly common view in medicine where anatomy texts took masculine anatomy and symptoms to be standard. We also find this in Freud's theory of human development where females were thought to have the additional problem of coming to see that they were in fact malfunctioning males. We don't regard females are malfunctioning males anymore, however. Instead we regard xx (female) and xy (male) to be dif-functions in the sense that they are two different (though equally functioning) ways of being. If someone is xxy or xxx, however, then we typically regard them to be malfunctioning. There is a current movement for these alternative genotypes to be regarded as dif-functions rather than dysfunctions, however. One biologist has argued that there are 5 sexes rather than 2 and it is hard to see how purely causal facts

can settle these issues one way or another. The relevant issue is how purely causal historical processes determine whether something is a function or a dysfunction or a dif-function.

Murphy maintains that the two-stage view is committed to the idea that functions and malfunctions are independent of human interests.

> ...on the two-stage view, the criteria for assessing adequate performance are supplied by nature rather than by a human practice... It is not the view that relative to human goals and interests, we can establish what psychological systems should be like and how they should be arranged to meet those goals and further those interests. Rather, it is the view that psychological normality imposes non human, natural functional standards. Those standards exist independently of what people think they should be.

He doesn't think that this notion of function creates a problem with respect to non- normative causal processes on the one hand and a normative notion of function and malfunction on the other, however.

> "Some people will say that since even this view licenses statements about what some biological system ought to be like, it is in fact normative in a fairly weak sense... All of medicine is normative in this sense – the problem is whether any science is not, though, because all sciences license expectations about what ought to happen in a normal system: stars, for example follow a reliable progression through developmental stages, so we can predict what ought to happen to them'. P 85

In the literature on disorder there has been a sustained critique of the notion that function can be used to ground psychiatry in non-normative facts where those non- normative facts consist in purely causal processes. The main line of argument comes from those who think that function of parts are determined by teleological functions of the whole. This line of criticism is largely inspired by Aristotle's teleological notion of function and it seems to rely on Aristotelian notions of a 'good person'. While I don't wish to

engage with Aristotle's view here I do think that there is something to this line of criticism. We can see the problem as one of a complete description of the causal processes in our world failing to entail facts about function and malfunction. This is the familiar point that there seems to be a gap between purely causal processes on the one hand and normative facts about function and malfunction on the other.

With respect to Murphy's example of the stars developmental stages it would seem to me that models of **reliable** progression are models of statistically normal or average progression. If a star does not progress through the stages that the model describes because of intervening causes then I don't think that we describe the star as malfunctioning except insofar as malfunctioning is analysed as deviation from the norm or average that we have built into our model. Mozart was statistically abnormal or deviant with respect to his musical abilities but we don't usually want to say that he was malfunctioning in virtue of his statistical deviation. Similarly, if we attempt to say that xxx is a dysfunction because it is statistically abnormal this doesn't seem to be a satisfactory analysis of the relevant notion of dysfunction. It seems plausible that entire populations could be malfunctioning in the sense of having some medical condition like broken legs or infestation by parasites and it similarly seems plausible that mental disorders could turn out to be far more prevalent than we had supposed. The statistical notion of abnormality thus does not seem to be the relevant notion for an explication of the bio-medical notion of dysfunction. The statistical notion does seem to be the relevant notion in the case of the star, however. One thing that I find interesting here is that insofar as we can say that the star is malfunctioning it is malfunctioning in virtue of falling short of the statistically average process that we have built into our model.

Now, we might have intuitions that these alternative genotypes are appropriately regarded as dysfunctions rather than dif-functions in virtue of the individuals with those genotypes being harmed by them. There seem to be cases where individuals are harmed by things that aren't malfunctions, however, and we do not regard all harms to be mental or physical disorders and

so we would need to know more about the relevant notion of harm. One might maintain that these alternative genotypes are malfunctions because the individuals are unable to reproduce. If we consider things from the level of group selection rather than individual selection and individuals with other sexes were found to invest heavily in their kin, for example, then it isn't obvious that they are malfunctioning compared with dif-functioning, however. It might be the case that there is a fact of the matter as to whether individual selection or group selection is the relevant process for fixing the functions and malfunctions but this doesn't seem obvious to me.

Wakefield maintains that there is something special about natural selection with respect to fixing 'natural functions' that obtain independently from us. When he attempts to explain what is special about natural selection he appeals to our explanatory interests, however. **If** we are interested in knowing what it was that past tokens did that accounts for their survival and reproduction **then** evolution by natural selection is the relevant causal process. We first need to identify survival and reproduction as the effects that are relevant for fixing the functions, however. The next step is to identify the effects of mechanisms where those effects contribute towards survival and reproduction or away from survival and reproduction. The notion is (roughly) that if an effect contributes towards survival and reproduction then we have grounds for considering the effect to be a function of the mechanism whereas if the effect hinders survival and reproduction then we have grounds for considering the effect to be a malfunction of the mechanism. This is, of course, a very rough picture. There are issues to do with whether causal processes are enough to fix functions or whether we need to invoke counter-factuals as well. I'm not attempting to offer necessary and sufficient conditions for natural function here, however, I'm just trying to very roughly convey the line that Wakefield and others are trying to run. What is important to note is that the identification of survival and reproduction as the relevant standard seems comparable to the identification of statistical average as the relevant standard in Murphy's example of the star.

It seems that while one can't get normativity from purely causal processes

one might be able to get normativity from a conjunction of our explanatory interests together with non-evaluative facts about statistical averages or causal processes. This conclusion is less disturbing for a science of psychiatry than the conclusion that mental disorder is determined by our moral or social evaluations as the anti- psychiatrists maintained, however.

Murphy maintains that it is far from obvious that the relevant notion of function to ground psychiatry in causal facts is the evolutionary notion of function. In particular, Murphy and Woolfolk maintain that it seems possible that mental disorders could result from harmful failures of spandrels, or ex-aptations. One example could be that if the mechanisms that subserve language don't have the evolutionary function of enabling us to read this wouldn't undermine the status of dyslexia as a disorder. Murphy makes a case for science modelling Cummins functions rather than evolutionary functions in some instances and he maintains that Cummins functions seem more relevant for the medical sciences than the evolutionary notion of function. The notion of a Cummins function is the sense of function in which it is true to say that Harvey understood the function of the heart centuries before Darwin. It seems that Cummins notion of function may be more relevant for the medical notion of disorder.

Attributing a Cummins function to some mechanism (such as a heart valve) seems to similarly require us to identify or choose some output of the overall system that fixes the function of the parts, however. If we grant that the relevant effect of the heart is the pumping of blood then we can attribute functions to the parts of the heart with respect to what contribution they make to the hearts pumping blood. If we want to say why the function of the heart is to pump blood then we can appeal to the role that the heart plays with respect to the biological homeostasis of the organism (or something along those lines). The problem then becomes how we identify the biological homeostasis or survival as the relevant function of the organism. It seems that Cummins functions aren't able to ground function and malfunction in objective facts as we are required to identify what it is that the overall system is 'supposed' to do before we read off functions and malfunctions of the parts

relative to what it is that we think the overall system to be 'supposed' to be doing.

Murphy maintains that the malfunction assumption does for psychiatry what the adaptationist assumption does for evolutionary biology. He goes on 'which is to say that sometimes the malfunction assumption is false, sometimes we don't know whether it is true or false but that does not impugn diagnosis'. One thing that concerns me about the malfunction assumption, however, is that it is supposed to be what grounds psychiatry as a non-evaluative science and that it seems to recommend a methodology for modelling mental disorders. The methodology seems to be that we model 'normal' or 'functional' biological or psychological processes and then we explain disorders by appealing to breakdowns in the model. Much work in the cognitive neuro-sciences and the bio-medical sciences has been done utilising this approach. We have explanations that characterise delusions as being the result of some kind of breakdown in belief formation and / or retention mechanisms; we have explanations of autism as a theory of mind deficit and so forth. The malfunction assumption can't make much sense of other projects that have been done, however. Instead of working with the malfunction assumption some theorists have worked with a function or adaptationist assumption where certain traits (such as histrionic or psychopathic) may be modelled as evolutionary adaptive strategies. Some theorists have attempted to characterise disorders such as depression, schizophrenia, and anxiety as evolutionary adaptive strategies that result in harm in present environments because environmental circumstances are far removed from those in savannah life.

While I'm not going to look at the plausibility of particular theories that have been offered my main point here is that the malfunction assumption does not seem to be required in order for us to study mental disorders scientifically. Instead of attempting to model mental disorders as deviations from some standard one could simply describe the causal processes that seem relevant for some behavioural output while remaining neutral on whether that behavioural output is adaptive or maladaptive. Science can thus model the

causes of certain kinds of behavioural symptoms even in the absence of the malfunction assumption. What seems harder to do in the absence of the malfunction assumption, however, is to say what it is about certain conditions or people that means that they are disordered.

My last objection to Wakefield is something that I shall just touch on briefly though it is something that I want to develop. The objection is that he commits himself to too much a-priori when he maintains that malfunction is necessary for mental disorder (and in particular when he identifies causal-historical processes as being necessary for mental disorder). A lot of theorists have attempted to construct counter-examples where we intuitively regard the person to have a mental disorder and yet where it is stipulated that they do not have an inner malfunction. Wakefield then responds to these objections by maintaining that either there is malfunction after all (and thus the alleged counter-examples actually provide support for his view) or that he is not inclined to regard the individual to be mentally disordered since there is no failure of inner function. Wakefield thus maintains that our intuitions about who is and who is not mentally disordered should be revised to be in keeping with the malfunction assumption. His critics maintain that conversely our intuitions about malfunction should be revised to be in keeping with our intuitions about which individuals are and are not mentally disordered. There seems to be a bit of a stand-off with this tactic.

If we take a step back from the debate it seems to me that what is going on here is that we have three main intuitions about mental disorder.

- - The first intuition (or set of intuitions) are around which people are appropriately regarded as mentally disordered and which conditions are appropriately regarded as mental disorders. This intuition is important with respect to the role of prototypical cases helping us fix the reference for our term.
- - The second intuition is that mental disorder is a natural kind term. We think that science will discover whatever it is that the prototypical cases have in common that people without mental disorders lack.

- - The third intuition is the dysfunction assumption or the notion that people with a mental disorder have an inner malfunction. This basically captures our intuition that there is something wrong with these people.

In maintaining that mental disorder involves inner dysfunction a-priori Wakefield makes the third condition essential and thus non-negotiable. There seems to be a tension between his maintaining that the relevant dysfunctions are to be determined by science on the one hand and his stipulating that science must discover a relevant dysfunction on the other. If scientists succeeded in offering an evolutionary adaptive account of mental disorder, however, and did not characterise them as the result of inner dysfunctions then Wakefield would be left having to conclude that there aren't any mental disorders. Some anti-psychiatrists maintain that mental disorders do require inner malfunction then go on to argue that prototypical cases of mental disorder do not involve inner malfunction. They thus seem to agree that inner malfunction is necessary for mental disorder and their disagreement comes down to whether scientists will discover that the prototypical cases of mental disorder have an inner malfunction or not. I think that it would be unwise to make any of the above intuitions essential to an account of mental disorder. While they are strong intuitions that go some way towards helping us fix the reference it might be that we need to revise our assumptions depending on how the world turns out. If the facts about function and malfunction can't be read off purely causal facts then it would be hard to see what sense to make of the notion that scientists 'discover' functions and malfunctions by investigating purely causal processes, however.

A problem remains with respect to defining mental disorder. If malfunction isn't necessary for mental disorder or if malfunction isn't a matter for science to discover then what is it that grounds psychiatry or medicine as a scientific discipline? The anti-psychiatrists often maintain that prototypical cases of mental disorder don't involve inner malfunction so much as their behaviour violating certain kinds of social or moral norms. They don't say much more about what kinds of social or moral norm violation are relevant, however, and it seems clear that there are many kinds of social and moral norm violations

(such as laziness or strangeness or moral 'badness') where we don't regard the person as being mentally ill. Without more of an account of what it is about their behaviour that we regard to be indicative of disorder their view seems implausible as it stands.

While survival and reproduction are fairly obvious standards for fixing functions if we are interested in evolutionary biology it does seem that our explanatory interests play a crucial role in allowing us to get normativity from the notion of function and malfunction. With respect to medicine there is widespread agreement that disorders that threaten a persons survival are disorders and the reasonableness of this view seems to be inherited from the reasonableness of survival as something that we are interested in promoting. Despite widespread agreement that some conditions are disorders there are controversial cases in medicine, however. There are some conditions where it is unclear whether they are disorders or mere problems in living or whether surgery is a medical requirement or merely elective. The further away one gets from issues of survival and the more expensive the treatment the more controversy there is as to the status of the individual or the condition as appropriately being regarded as disordered. Psychiatry doesn't seem to be concerned with survival of persons in quite the way that biology and medicine are.

So in my talk today my conclusion is largely negative. The claim is that the malfunction assumption can't do the work that is required of it. The biological notion of function can't ground psychiatry in facts about purely causal processes because the biological notion of function requires us to identify survival and reproduction as the relevant features for fixing functions. Given the explanatory interests of evolutionary biology this is a reasonable thing to do and I have no problem with the scientific status of biology. It is unclear, however, that survival and reproduction are the relevant features for fixing functions and malfunctions for psychiatry. I think that much more work needs to be done on the harm component with respect to understanding what disorders have in common. Coopers has stated that mental disorders might be a little like the notion of 'weeds'. That is to say that our values

might well be crucial for determining the class of things that we are interested in. While there is no objective science of weeds because weeds don't have non- evaluative properties in common that differentiate them from non-weeds there can still be a scientific classification of plants, however. It could similarly be the case that individuals with certain kinds of mental disorders share certain causal processes in common though our values are an important part of how we distinguish the mentally disordered from the non-mentally disordered. While the anti-psychiatrists maintain that the relevant values are social or moral it is not the case that any social or moral norm violation is indicative of mental disorder. More work needs to be done on what kinds of norms are relevant. The notion of malfunction isn't very explanatory if it is merely an assumption that we have built into our model where we could re-describe the casual processes that we have discovered as dif-functions instead of dysfunctions.

Chapter 3

Towards a scientific nosology of psychiatric disorder

Presented to the Philosophical Society of the Australian National University

Introduction

What I'm hoping to do in this talk is to provide an introduction to some issues in the philosophy of psychiatry in order to motivate and help clarify the scope of my thesis. Philosophy of psychiatry is a fairly broad topic and there are a number of different aspects that one could write a thesis on. I'll start by saying a bit about what I'm not going to talk about and then I'll get on to some of the issues that I want to investigate in more depth in my thesis. Firstly, I'm not going to talk about issues of autonomy or issues of moral or legal responsibility for ones actions. I'm also not going to talk about the nature of evil, or the extensive literature that has accumulated on psychodynamic theories of functioning and disorder. Lastly I'm not going to talk about what it is like to be mentally ill in the sense of reading case studies to grasp the phenomenology of it all.

What I do wish to focus on is a cluster of issues in psychiatric nosology.

More in particular I want to focus on what makes a good taxonomy a good taxonomy and more in particular how psychiatry can progress as a science with respect to the development of a more adequate classification system. Such a project is part of the philosophy of science and related areas in the empirical philosophy of mind. As a science of the mind psychiatry needs to develop in line with the cognitive neuro- sciences and thus philosophical work that has been done on integrating these two disciplines will be relevant to my thesis project.

Part of what I want to do in my thesis is thus to consider how the genetic, neurological, cognitive, behavioural, social, and environmental facts relate to one another. I'm hoping to develop a philosophical framework for an integrative approach to the science of psychiatry where facts from each of those levels could well turn out to be relevant for psychiatric nosology and models of different kinds of disorder may well need to incorporate facts from more than one of those levels.

Dominic Murphy in his book *'Psychiatry in the scientific image'* has recently stated that in order to progress as a science psychiatry needs to move beyond purely behavioural symptoms and look to the cognitive neurosciences for the causal mechanisms that sustain the behavioural symptoms of psychiatric disorder. I agree with him in this, but I think that not all such causal mechanisms are internal to the agent. While Murphy does consider the rule of social causal mechanisms I think that there is a lot more work to be done on this.

A rough outline of what I want to get through today is as follows: Firstly, I'm going to start by introducing the two main systems of classification; the ICD index and the DSM IV. Then I'll consider the purpose of psychiatric nosology so we are better able to assess whether a system is adequate. The last issue I want to consider is the different kinds of categories that mental disorders could turn out to be and some implications of our finding out that a kind of mental disorder is mostly this or that kind of category and in particular whether social causal mechanisms can result in categories that are viable for scientific investigation

Two Psychiatric Nosologies: The ICD and the DSM

There are two main nosologies or systems of classification of mental disorders. One of these is the *International Classification of Diseases Index*, otherwise known as the ICD. The ICD was developed by the World Health Organization for the purpose of compiling statistics on the prevalence of a variety of medical causes of death. As such psychiatric disorders comprise one section and the other sections are constituted by a variety of other medical conditions that are not regarded as psychiatric. The current edition of the ICD index is the ICD-10, or the 10th edition.

The other main classification system is the *Diagnostic and Statistical Manual of Mental Disorders*, otherwise known as the DSM. The DSM was developed by the American Psychiatric Association and it focuses solely on classifying psychiatric disorders. As such it is more specific than the ICD index and it contains more kinds of categories. There have been a number of editions and text revisions in which adjustments are made. The current edition of the DSM is the DSM-IV-TR, or the text revision version of the fourth edition. The DSM states three main aims: Firstly, to compile statistics on the prevalence of different kinds of disorder. Secondly, to facilitate research on mental disorder. Thirdly, to facilitate communication between clinicians. I shall return to how these aims interrelate shortly.

While the use of the DSM as a system of classification was largely restricted to the United States of America initially, its use has increased around the world such that the majority of clinician's now provide DSM diagnostic codes when they are classifying individuals. The development of translation manuals that allow clinician's to translate a diagnostic category and code for a mental disorder from one system of classification into another has assisted with this. In the United States of America I'm fairly sure that both ICD and DSM codes are required for health insurance purposes.

Identifying Psychiatric Disorders

Diagnosis of mental disorder seems to consist of (as least) two interrelated components. Firstly there is the issue of how we identify whether or not an individual is mentally disordered, and secondly there is the issue of how we identify what particular kind of mental disorder they have. I shall address both of these in turn. With respect to the first issue of identifying mental disorder in general we can distinguish two further related problems. The first is how to distinguish a disorder from a 'problem in living', The second is the issue of how to distinguish mental or psychiatric disorders from non-mental, neurological disorders, or general medical conditions.

With respect to the first issue the DSM provides a global assessment of functioning (or GAF) scale that is meant to capture the extent of the disability, disorder, dysfunction, or distress. Without significant impairment in functioning a clinician should not diagnose an individual as having a mental disorder even if they meet diagnostic criteria for a particular kind of mental disorder. The GAF scale reflects the notion that the DSM is primarily concerned with providing a tool to enable clinician's to make diagnostic decisions. The DSM also lists the following features that clinician's are supposed to use to assess whether an individual has a mental disorder: statistical infrequency, violation of norms, personal distress, disability or dysfunction, and unexpectedness. With respect to unexpectedness Davison and Neale (p.6) maintain that, for example, 'an anxiety disorder is diagnosed when anxiety is unexpected and out of proportion to the situation, as when a person who is well off worries constantly about his financial situation'. The DSM takes this list to not only be a way of identifying individuals on whom to intervene, however, it takes it as an attempted definition of the nature of mental disorder, though it is acknowledged that current definitions are inadequate to capture the phenomenon that is of interest.

The most influential definition of disorder is probably Wakefield's 'Harmful Dysfunction' (HD) analysis of the concept of disorder as it is employed in psychiatry, medicine, and common sense. Wakefield maintains that there are

two individually necessary and jointly sufficient conditions for someone having a disease, disorder, or illness. The first condition is that there is an inner malfunction and the second condition is that the effects of the inner malfunction are harmful to the person and / or to society. Wakefield maintains that a clinician is justified in classifying an individual as mentally disordered when a clinician believes that their harmful behaviour is the result of inner malfunction. Wakefield's account is controversial, however. While he takes the notion of malfunction to be determined by facts about biological function as talked about by theorists such as Millikan and Neander, other theorists have denied that there are facts about the person that determine that an individual has a disorder. Instead, they maintain that the relevant notion of malfunction is dependent on our value judgements that the individual's behaviour is in violation of norms of society. Even if we grant that there is more to disorder than social norms another point of controversy is over whether disorders must be due to inner malfunction as opposed to outer malfunction, or problems relating to their society or environment. These are issues that I shall return to in a later section.

With respect to the second issue of distinguishing psychiatric disorders from neurological disorders at a first pass mental disorders might be thought of as disorders of cognitive processes, such as thinking, emotion, or desire. Current classification regards cortical blindness as neurological rather than psychiatric, however. This move seems to be in line with our common-sense intuitions though it is in tension with our intuition that mental disorders are disorders of cognitive processes as vision would be a paradigmatically cognitive or mental process. Indeed, other visual disturbances such as hysterical blindness and hallucinations are typically regarded as psychiatric rather than neurological. The concept of mental that is employed in both common sense and in current nosology thus seems to be under-inclusive. Current nosology might also be thought of as over-inclusive, however. For instance, the essential feature of Tourette's is tics but there wouldn't seem to be anything particularly mental or cognitive about a motor disturbance. Perhaps Tourette's really has an essentially cognitive component that is neglected by

current nosology, or perhaps Tourette's is not appropriately classified as a mental disorder and current nosology is over-inclusive with respect to this case. It might be that the distinction between neurological and psychiatric conditions is nothing more than a historical relic of the type of intervention once thought appropriate where psychiatric disorders are treated by therapy and neurological disorders are treated by physical intervention. It might be the case that there is no principled distinction to be made between the subject matter of neuroscience or the sciences of disorders of the mind and the subject matter of psychiatry.

The ICD and the DSM are similar in the way that they distinguish between different kinds of disorders even though there are differences in the kinds of disorders that are provided by each classification system. They both provide clusters of behavioural symptoms, or cognitive symptoms identifiable by verbal behaviour. When an individual has significant impairment in their functioning and they meet enough of the behavioural symptoms then the person may be regarded as having that particular kind of mental disorder. While some of the kinds of disorder have essential symptoms the majority do not, rather the person only need exhibit a certain number of symptoms. There are also exclusion criteria such that when an individual meets diagnostic criteria for more than one kind of disorder one diagnosis may take priority and exclude the other. There are other exclusion criteria as well, such as that the behaviour isn't caused by a general medical condition or the effects of a substance or toxin, or that the behaviour isn't performed solely as a matter of political protest or religious conviction.

Ian Hacking maintains that even more important than the DSM definition of mental disorder and kinds of mental disorder the accompanying casebook that provides case studies of people who are prototypical instances of someone both being mentally disordered and meeting a certain diagnostic category. Clinical judgement may thus be thought to consist largely of experience with a variety of more or less prototypical cases so that a clinician's judgement falls in line with the judgement of other health professionals.

The Purpose of Psychiatric Nosology

The ICD classification system was developed so that statistics of the prevalence of the conditions that are leading causes of deaths could be compiled. The DSM states three related objectives in a nosology for psychiatric disorders, however. The first objective is to provide a system that facilitates research on psychiatric disorder. The majority of research takes the diagnostic categories provided by the DSM as the basic unit of research analysis. When people search for a genetic basis, the structural or functional neurological abnormalities, the efficacy of medication or therapy, the cross-cultural variation, or the course of illness, the DSM criteria is used to identify the individuals with the disorder that is the subject of research. While it is important to distinguish clearly between the nature of disorder on the one hand and how we go about identifying individuals with the disorder on the other, the two are clearly related in the sense that we need to identify individuals in order to commence investigation into the generalisations and predictions that we can make about them as a group and our findings about individuals in the group could lead to subsequent revisions of the diagnostic categories.

The relevant notion here is the notion of construct validity. The DSM provides a list of constructs, or kinds of disorder. A construct is thought to be valid when there are scientific generalisations and predictions that can be made about an individual on the basis of identifying the individual as an instance of the category picked out by the construct. As such, constructs can be more or less valid depending on whether they support more or less generalisations and predictions. The notion of a category that is in play here seems to be in line with Boyd's homeostatic property cluster theory of kinds where we note that there are observable properties (in this case behavioural symptoms) that are found to be clustered together in nature. Because these properties are found to be clustered together we can form a construct of the category and we can make fairly accurate generalisations from the presence of some properties, or symptoms, to the likely presence of some other properties, or symptoms. When we observe some of those properties, or symptoms we can also make fairly accurate predictions such as response to treatment

or the future course of illness, for example. The homeostatic property cluster view might only be one way in which we could get projectability, however.

The DSM states that its third aim is to provide a classification system that facilitates communication between clinicians. Prior to the development of the DSM and ICD index there were a proliferation of nosologies that were very theory dependent on which variety of psychodynamic theory the theorist subscribed to. Part of the motivation from moving from a classification of inner causes to a classification of behavioural symptoms was that regardless of theoretical orientation clinician's could agree as to whether an individual exhibited this or that symptom. As the diagnostic categories are built out of behavioural symptoms this also allowed clinician's to agree as to what diagnosis a patient should have, regardless of the clinician's theoretical orientation. The issue here is thus one of inter-rater reliability. When a behavioural symptom or a diagnostic category has good inter-rater reliability then different clinician's would attribute the same symptoms and diagnostic category to the same individual. Both construct validity and inter-rater reliability would seem to be required in order for compilation of statistics on prevalence rates to be meaningful.

These two aims of facilitating research and promoting communication between clinicians might be thought to map onto two different aims of providing a nosology that is scientifically fruitful with respect to generalisation and prediction and providing a nosology that is useful for clinicians with respect to identifying which individuals are requiring intervention. The DSM takes these aims to be complimentary and indeed they do seem to be related. One would hope that nosology is useful with respect to identifying what kind of disorder an individual actually has, for example, and one would also hope that a scientific nosology would provide information as to what kinds of interventions are likely to be effective. It might turn out to be the case that these aims diverge, however. While purely behavioural symptoms might be most useful with respect to identifying the individuals who require intervention purely behavioural symptoms might be less than optimal with respect to enabling us to identify the underlying causal mechanisms that provide

information as to the optimal points of intervention.

While the GAF scale seems practically important in that it is focused on whether intervention is called for it is not a consideration that should guide scientific study on causal mechanisms and intervention points relevant to it. One issue that I want to deal with is what nosological categories would have to be like in order for a scientific nosology to be possible. I'll now turn to considering some of the different kinds of categories that mental disorders could turn out to be.

Kinds of Categories: Part One

Essential Kinds are thought to be categories that share the same intrinsic, or non- relational essential properties. Paradigmatic examples include water and gold where in order to count as an instance of water the instance must have the property of being H2O and in order to count as an instance of gold the instance must have the property of being atomic number 79. The intrinsic properties are thought to be constitutive of kind membership. Mental disorders could turn out to be essential kinds if it was found that they had a very specific biochemical basis, for example.

Biological Kinds. Are thought to categories that share the same relational, extrinsic essential properties of historical lines of descent. Paradigmatic examples include elms and tigers. There is controversy over whether natural kinds are required to have intrinsic essential properties such that biological kinds don't count as natural kinds; or whether biological kinds are natural kinds and thus natural kinds may have extrinsic, relational essences; or whether membership of a lineage is an internal property to the species as a whole and thus biological kinds are intrinsic essential kinds and thus natural kinds after all. I shan't get caught up in this debate, however. Whether biological kinds are properly thought of as natural kinds or not it seems that they form something of a natural category.

The notion of a natural category is tied up with the notions of generalis-

abilitly, projectability, and predictive leverage. Natural categories may be thought of as something along the lines of what Boyd calls a homeostatic property cluster. The notion here is that certain properties are found to be clustered together in nature. If we see some properties then we can infer the presence of other properties and thus homeostatic property clusters support scientific generalisations and predictions. We would seem to identify instances of a natural category on the basis of these observable properties. The category of birds, for example, includes such properties as flight and feathers where these properties are superficial properties rather than properties at a lower level of analysis such as genetic. This view seems to be very much in line with the way the DSM provides behavioural symptoms as relatively superficial observable properties that enable clinicians to identify individuals as having a certain kind of disorder. The majority of diagnoses also do not have essential symptoms and thus members of diagnostic categories exhibit family resemblances of symptoms. A feature of the property cluster view is that different instances have slightly different features and they may be more or less prototypical, for example, not all birds can fly.

While the DSM provides a nosology where clinicians identify mental disorder on the basis of behavioural symptoms it would seem to be a separate issue whether mental disorders are constituted or defined by the behavioural symptoms as **Behavioural Kinds**, however. If one takes the behavioural symptoms to be definitional or constituitive then there could plausibly be borderline cases where it is indeterminate whether the individual is in fact a member of the kind or not. It would seem, however, that the main reason why it is that certain properties are to be found clustered together in nature is because they share some underlying causal mechanism that are responsible for the properties homeostasis. It is because the causal mechanism is found in the different instances that we are able to make scientific generalisations and predictions. It could also turn out that the same set of behavioural symptoms could be generated in two quite different ways. If we found this to be the case then it would seem better to conclude that there are two distinct kinds of disorders where different interventions are required.

Thus, while we might typically identify or come to believe that instances are members of a certain category on the basis of superficial, observable properties, taxonomy is often revised as we come to define categories on the basis of the underlying causal mechanisms that are necessary for category membership. This is because causal mechanisms seem to be what leads to the properties homeostasis and the more homeostatic a property cluster the more those properties are able to support generalisations and predictions. While Boyd's view focused on internal generative mechanisms it is unclear whether a principled distinction between internal and external generative mechanisms can be sustained. If one views a species as an individual, for example, then lineage would be an internal property to the species. Boyd's homeostatic property cluster view, or something like it, can thus be thought of as consistent with both the essentialist and relational view of categories.

Wakefield attempts to draw a principled distinction between the 'right' and 'wrong' kind of causes for mental disorder. He maintains that when the harmful behaviours are due to inner malfunction the individual is mentally disordered and when the harmful behaviours are the result of external causal mechanisms the harmful behaviours are not indicative of mental disorder and are instead best thought of as a non-pathological problem in living. It would seem that whether mental disorders are constituted by social causal mechanisms would be an empirical matter rather than one to be settled on intuitive grounds or by stipulation, however. Wakefield is especially focused on the notion of neurological and / or cognitive malfunction which he characterises along the lines of a hardware / software distinction and while he doesn't mention it I don't think he would be opposed to adding genetic malfunction to the mix (supposing that it makes sense to talk of genetic malfunction or kinds of genetic disorder). This way of thinking about inner malfunction seems very much in line with cognitive neuropsychology and it might be the case that the kinds of psychiatric disorder are derived as malfunctions of the causal mechanisms that is identified, at lest in part, by cognitive neuroscientists. **Neurological kinds** would seem to be fairly straightforwardly thought of as biological kinds. Some theorists have attempted to analyse

Psychological kinds as another variety of biological kinds where mental or cognitive states such as belief and desire are the kind of state they are in virtue of what the mechanisms that support the state have evolved to do.

Sometimes theorists (like Wakefield) appeal to current functions instead of evolutionary functions where the effects of a current function are responsible for the mechanism being prevalent in current populations. Treating mental kinds as biological kinds is controversial, however. The natural categories or kinds would seem to be those of normal functions. Psychiatric kinds are breakdowns of normal symptoms and the breakdowns may be unified only by being breakdowns of a specific mechanism. There would thus seem to be an open ended class of ways things could go wrong. Attempting to list them all with respect to behavioural symptoms is thus bound to get unwieldy and more progress might be made by looking at different ways that normally functioning systems can break down. I now want to turn to some of the external mechanisms that might be relevant for mental disorder and I'll consider several different varieties of socially constructed kinds.

Kinds of Categories: Part Two

Artefacts like pens and chairs are paradigmatic examples of **Socially Constructed Kinds**. Instances of the category pens count as members of the category in virtue of having the historical relational property of being designed by an agent for a certain function. As such agents designing them for a certain function is necessary and sufficient for or constitutive of category membership. Because they are designed by agents for a certain function pens exhibit a cluster of superficial properties in common. Those properties may enable us to identify instances as instances of the category. If we found something that shared the superficial properties with pens but it grew on a tree or materialised out of a swamp then because it was not designed by an agent with the relevant intention it would not count as a pen, however. While pens are dependent on us for their initial existence once the instances have been brought into being then it is a mind independent fact that the

instances are in fact members of the category. Even if we lost our concept of a pen or we no longer used pens to perform their function the instances that still exist would continue to exist as members of the category.

Some other socially constructed kinds aren't dependent on the intentions or mental states of agents so much as their social practices. Something might count as a doorstop, for example, not because it was designed with that intention in mind, but instead because it is currently being used to perform that function. If we accept this reading of what it is to count as a doorstop then it would follow that if we were to stop using the object as a doorstop that it would cease to be a member of that kind. There isn't a science of pens or doorknobs. While we might be able to make generalisations such as that pens usually have ink and that doorstops tend to be sturdy or obstructive it would seem that there are significantly less generalisations and predictions available to us than there is with either chemical or biological kinds.

I now want to turn to another sort of socially constructed kind that is clearly more relevant to psychiatric disorder. The notion of a **Looping Kind** was initially introduced by Hacking and it has subsequently been picked up on by other authors such as Griffiths, Mallon, and Murphy. In order to describe the features of looping kinds I need to draw a further distinction between what I shall call explicit looping kinds and implicit looping kinds.

Explicit looping kinds are kinds that are constituted by our social practices. While artefacts like pens are mind independent in the sense that they continue to be pens in the absence of our social practices around them, looping kinds are thought to be causally rather than definitionally or constitutively dependent on our social practices. Our social practices cause them to come into being as instances of the category and if our social practices change then this can cause them to go out of being as instances of the category. It is easiest to see this by way of example. Members of Parliament and Licensed Dog Owners are examples of explicit looping kinds. We have social practices around parliament and the election of members of parliament, for example, and in virtue of those social practices individuals come to be Members of Parliament. Unlike pens explicit looping kinds aren't independent of our

social practices because if we alter our social practices so that there isn't a parliament then the individuals would cease to be members of the category Members of Parliament.

Individuals that are Members of Parliament have properties in common such that they may be identified as Members of Parliament. We are able to make generalisations and predictions about Members of Parliament with respect to the properties they exhibit or are likely to exhibit and ways in which they are likely to behave. When the individuals are no longer members of the category Members of Parliament then they lose the properties that they had in virtue of their category membership, however, and we can no longer make such generalisations and predictions about them. These looping kinds are explicit in the sense that we are aware that the categories are dependent on our social practices. We know that there wouldn't be any Members of Parliament if we altered our social practices in certain ways. This doesn't stop us being able to make generalisations and predictions about Members of Parliament, however. It also doesn't stop the special science of politics from taking them seriously as a category.

Implicit looping kinds are similar to explicit looping kinds except that in this instance we aren't explicitly aware that the instances of the category are instances of the category because of our social practices and instead we regard the category as being a natural (or biological) kind. Hacking maintains that in this case if we were to become aware of their status as a looping kind then it would be inevitable that our social practices would change and this would have the result that the instances would no longer being members of the category. Our awareness and subsequent change in our social practices would also result in an alteration to the properties that the individuals shared as members of the category and thus the generalisations and predictions that were made about individuals in virtue of their category membership would no longer obtain.

Once again, it is probably best to convey this phenomena by way of example. Examples of implicit looping kinds include categories such as demonic possession and being possessed by a wild pig. The notion is that when we

believed in these concepts then our belief in them and our social practices around them results in opening up new ways of behaving that are stereotypic of the category. If we take a person to be a member of the category or if they take themselves to be a member of the category then this may cause them to behave in ways that are stereotypic of the category. Members of the category are thus able to be identified as members of the category in virtue of sharing certain stereotypical properties in common. What is supposed to be distinctive about these categories, however, is that they cannot survive our realisation that they refer to looping kinds. The notion is that once we become aware that the properties are due to our social practices then we cease believing in them and we inevitably alter our social practices so that the individuals no longer display those common features.

This phenomena is probably best conveyed by way of Ian Hacking's characterisation of Multiple Personality Disorder which he takes to be an 'all too perfect illustration of the feedback effect' in implicit looping kinds:

> We tend to behave in ways that are expected of us, especially by authority figures – doctors, for example. Some physicians had multiples among their patients in the 1840's, but their picture of the disorder was very different from the one that is common in the 1990's. The doctors' vision was different because the patients were different; but the patients were different because the doctors' expectations were different. That is an example of a very general phenomenon: the looping effect of human kinds. People classified in a certain way tend to conform to or grow into the ways that they are described; but they also evolve in their own ways, so that the classifications and descriptions have to be constantly revised. (Hacking, 1995, p. 21).

Hacking thus maintains that in the case of implicit looping kinds there is a tension in that possession of the concept and our social practises around this are the mechanism that both stabilises and destabilises the property cluster. With respect to the stabilising function he considers that individuals symptoms are shaped because when the clinician applies the concept to the

patient this results in the clinician having either implicit or explicit expectations of the symptoms they expect to find in the patient. This changes the way that the clinician relates to the patient and is thought to lead to the patient exhibiting the symptoms they are expected to exhibit. Another way this can happen is if the clients apply the concept to themselves and thus come to exhibit symptoms that they believe to be stereotypic features of the category. In this way the concept and our social practices stabilise the symptoms that the patient exhibits as they come to behave in ways that are consistent with the stereotype.

Hacking also considers how our social practices can have a destabilising effect, however. He traces how the stereotypical features of Multiple Personality Disorder have evolved through time. Hacking tells a complex story of destabilisation and he draws on a variety of factors including political and theoretical, which lead to our beliefs about the concept evolving and the symptoms evolving in response to this. Some examples he has of this effect in the case of MPD include how many alters are thought to be typical (one or several or over one hundred); whether there is one or two way amnesia; how long it takes to switch between alters; and reports of abuse. It thus seems that the change seems mostly to be a function of a change in the theoretical views of clinicians. This led to a subsequent change in how they related to their clients and what kinds of symptoms they expected to see. Hacking seems to regard implicit looping kinds as having some homeostasis but the homeostasis is less stable than other kinds of socially constructed and natural kinds in that awareness of their status as looping kinds will result in the dissolution of the category.

Implications of Implicit Looping Kinds for a Scientific Nosology.

In these cases because it is implicit that we are dealing with a looping kind we are unaware of the impact of categorisation, our social practices, our expectations, our ways of interacting with the person, and so forth. If we

come to believe that a certain kind of mental disorder is a looping kind then it seems that one of three things could happen: Firstly, it could turn out to be the case as an empirical matter of fact our change in belief does not result in a change in our social practices. While Hacking thinks the relevant social practices are ones that invariably would change if we became aware that the category was a looping kind surely it could be possible that the social practices that are sustaining the phenomena could be resistant to change possibly because they have other beneficial effects. It is unclear whether Hacking would consider this to be an example of an implicit looping kind because it was implicit even though awareness did not result in its dissolution or whether Hacking would consider this to be an example of an explicit looping kind because it does not dissolve in the face of our awareness even though the so called explicit looping kind was implicit for a time.

Secondly, it could turn out to be the case that as an empirical matter of fact that if we came to believe the category was looping and we changed the relevant social practices the stereotypical behavioural features remain. In this case we seem to be left having to conclude that the category wasn't a looping kind after all. While it could still be socially constructed in the sense that artefacts similarly rely on us for their initial existence the phenomenon wouldn't seem to be dependent on our social practices and thus it would not be an implicit looping kind on Hacking's account. The third thing that could happen would be that our awareness of the category as an implicit looping kind could cause the stereotypic features to shift. If we found that a particular kind of mental disorder was an implicit looping kind this isn't to say that all instances of the category are suddenly cured of all symptoms of psychopathology, however. It is just to say that they won't display features of psychopathology that were stereotypic of the looping kind. They may well go on to display stereotypic features of another psychiatric kind, for example. Social constructionists about Multiple Personality Disorder often say that there is no such category as Multiple Personality Disorder there is only Borderline Personality Disorder that has been worked up into Multiple Personality Disorder in response to our social practices around the concept. The

notion here seems to be that if we refuse to participate in those social practices the patients will display stereotypic features of Borderline Personality Disorder instead.

What is unclear, however, is whether this would be so because the clinician's expect them to come to display the stereotypical features of Borderline Personality Disorder or whether this is in response to some other mechanism. If clinicians came to believe that there was no such category as Borderline Personality Disorder then would the individuals continue to behave in a way consistent with a diagnosis of Borderline Personality Disorder or would their behavioural symptoms shift so that they met criteria for another diagnostic category? While Multiple Personality Disorder is often one of the favourite categories of those who maintain that we need to look at social causal mechanisms it is unclear whether other, more paradigmatically biological psychiatric kinds could turn out to be looping kinds or to have a looping kind feature to their behavioural symptoms. It could turn out to be the case that mental disorder more generally has a significant looping kind component. If this was found to be the case then this would seem to have significant implications for both the project of how we identify mental disorders and the project of how we develop a scientific classification of them.

One implication is that focusing solely on behavioural symptoms might be counter- productive. Each subsequent edition of the DSM is praised for making scientific progress with respect to providing categories that better support generalisations and predictions. If the properties relevant for generalisation and prediction are purely behavioural symptoms and if the behavioural symptoms evolve over time in response to the classification system and a new round of expectations by clinician's then it would seem that the DSM approach will be limited insofar as the property cluster is unstable. The DSM may not only describe current symptamology but it also may have a causal role to play with respect to future symptom development. One consequence of this might be that the DSM and ICD aren't necessarily converging on constructs that are more valid than the old constructs; rather each edition might recover some of the construct validity that the old one had by

adequately capturing present symptoms that may, at least partly, have been evoked in response to previous systems of classification. Construct validity on the basis of generalisations and predictions on the basis of behavioural symptoms may be of limited value with respect to a scientific nosology.

If we identify kinds of mental disorders according to causal mechanisms rather than behavioural symptamology, however, then this enables us to say that the behavioural symptamology of a particular kind of disorder can evolve over time. This latter approach also allows that there could be considerable cross-cultural variation in the behavioural symptoms of individuals who have the same kind of mental disorder. While the DSM saw purely behavioural symptoms as progress from the causal mechanisms offered by the psychodynamic theorists cognitive neuropsychology would seem to have good prospects for grounding the next stage of scientific development from observational properties towards a scientific nosology of the causal mechanisms that produce psychiatric disorders. It seems plausible to me that more valid constructs may require us to incorporate causes from multiple levels of analysis. While there will be more to social causes than the looping effects that Hacking deals with the looping kind effect is interesting with respect to the relationship between social cognitive and behavioural facts. If we consider that the cognitive facts are represented within the brains of individuals it seems that whether the cause is inner or outer may be a function of how far back in the causal chain we look.

Chapter 4

The nature of mental disorder

Presented to the Australasian Association of Philosophy, New Zealand Division conference. Hosted by the Victoria University of Wellington

Presented to the Philosophy of Biology Graduate Student Workshop. Hosted by the Australian National University.

The Nature of Mental Disorder

There has been a lot of controversy over both the nature of mental disorders in general and also over the particular kinds of mental disorders that there are. The biggest threat to the prospects for a science of psychiatry is eliminativism where eliminativists maintain that we should eliminate our concept of mental disorder, as there is no such thing. One could also be an eliminativist about particular kinds of mental disorder that appear in psychiatric nosology. This variety of eliminativism would be less radical, however, as it is plainly the case that current nosology is a work in progress and it wouldn't seem to undermine the notion that there are categories of mental disorder it is just that we haven't hit upon them at present. There are a variety of ways that one could be led to eliminativism.

In this seminar I shall begin by talking about our concept of mental disorder and then turn to different ways our concept of mental disorder and concepts of particular kinds of mental disorder could turn out to map onto the world. There may be grounds for eliminativism here if our concepts don't map onto categories. I shall then turn to investigating different kinds of categories where there are different causal mechanisms that are responsible for generating the properties that make the category useful for scientific generalisation and prediction. It might be the case that social mechanisms are one such mechanism but that by their nature this results in categories that are less stable than the traditional sciences. If this turned out to be the whole story about mental disorder and there weren't any genetic, neurological, or cognitive causal mechanisms that were relevant then one might be led to eliminativism. If this is just part of the story, however, then the science of psychiatry would need to be reformed so as to investigate and better incorporate these social causal mechanisms.

The Concept of Mental Disorder

Defining mental disorder is problematic. The first problem is how mental disorders differ from other varieties of disorders. The second problem is how disorders differ from non-disorders. Another related issue is how we recognise or identify whether an individual is mentally disordered. I shall now address each of these issues in turn. With respect to the first problem, the issue is sometimes put as the problem of distinguishing mental or psychiatric disorders from non-mental or neurological disorders. At a first pass mental disorders might be thought of as disorders of cognitive processes, such as thinking, emotion, or desire. Current classification regards cortical blindness as neurological rather than psychiatric, however. This move seems to be in line with our common-sense intuitions though it is in tension with our intuition that mental disorders are disorders of cognitive processes as vision would be a paradigmatically cognitive or mental process.

Indeed, other visual disturbances such as hysterical blindness and halluci-

nations are typically regarded as psychiatric rather than neurological. The concept of mental that is employed in both common sense and in current nosology (or taxonomy) thus seems to be under- inclusive. Current nosology might also be thought of as over-inclusive, however. For instance, the essential feature of Tourette's is tics but there wouldn't seem to be anything particularly mental or cognitive about a motor disturbance. Perhaps Tourette's really has an essentially cognitive component that is neglected by current nosology, or perhaps Tourette's is not appropriately classified as a mental disorder and current nosology is over-inclusive with respect to this case.

With respect to the second problem the most influential view of disorder is probably Wakefield's 'Harmful Dysfunction' (HD) analysis of the concept of disorder as it is employed in psychiatry, medicine, and common sense. Wakefield maintains that there are two individually necessary and jointly sufficient conditions for someone having a disease, disorder, or illness. The first condition is that there is an inner malfunction and the second condition is that the effects of the inner malfunction are harmful to the person and / or to society. Wakefield's account is controversial, however. While he takes the notion of malfunction to be determined by facts about biological function as talked about by theorists such as Millikan and Neander other theorists have objected that the notion of malfunction is dependent on our value judgements that the individual's behaviour is in violation of norms of society. Another point of controversy is over whether disorders must be due to inner malfunction. These are issues that I shall return to in a later section.

If we now turn to the issue of how we recognise if an individual is mentally disordered it may well be that our concept of mental disorder is a cluster concept where there are several features that are relevant but where none of the features are necessary and where there is no clear boundary on how many features are sufficient for our regarding an individual to be mentally disordered. In textbooks on psychopathology the following features are commonly listed: statistical infrequency, violation of norms, personal distress, disability or dysfunction, and unexpectedness. With respect to unexpected-

ness Davison and Neale (p.6) maintain that 'for example, an anxiety disorder is diagnosed when anxiety is unexpected and out of proportion to the situation, as when a person who is well of worries constantly about his financial situation'. The Clinician's handbook The *Diagnostic and Statistical Manual of Mental Disorders* similarly attempts to define mental disorder by incorporating all of the above features and then conceding that no definition seems adequate to capture the phenomena. While Wakefield attempts to defend his 'Harmful Dysfunction' analysis arguing that it is an adequate analysis of our concept of disorder and that the DSM should revise its definition so it is in keeping with his account the DSM might be less interested in defining mental disorder and more interested in training clinician's to agree in how to classify individuals.

The DSM lists behavioural symptoms for each category and when a person meets a specified number of them they meet the diagnostic criteria for that disorder. Ian Hacking maintains that even more important than the DSM symptom lists is the accompanying case book that provides case studies of people who are prototypical instances of someone meeting a certain diagnostic category. Clinical judgement is thought to consist of experience with a variety of more or less prototypical cases so that a clinician's judgement falls in line with the judgement of other health professionals.

We also have our common sense conception of mental disorder. Our common sense conception seems to be similarly formed around exemplars of people who are considered to be mentally disordered where these exemplars form something of a prototype or stereotype. Prototypes seem to play a significant role in our intuitive judgements as to who is and who is not mentally disordered and also in what kind of mental disorder they have. Our common sense intuitions seem to have evolved as our conception becomes more informed by the categories of mental disorders offered by the DSM.

My main reason for dwelling on our conception of mental disorder is so we are in a better position to assess when we should be eliminativists about our concept of disorder when we see how the world turns out. I now want to say a few words about how our concepts can map onto categories and then

I'll turn to the main issue of this paper: the issue of the different kinds of categories that mental disorders could turn out to be. Along the way I'll consider varieties of reference and categories that could lead one to be an eliminativist about our concept of mental disorder in general or about our concept of a particular kind of mental disorder.

Different Kinds of Reference

I won't attempt to define a category at this stage as the notion should get clearer through this section and shall have much more to say about them in later sections.

The first variety of reference that I want to consider is **Nominal Reference**. When there is nominal reference a concept that is intended to refer to a category turns out not to refer to a category. Griffiths offers the example of Aristotle's notion of a SUPER-LUNARY OBJECT as an example of such a concept. The only property that the instances have in common is the property of falling under the concept and the instances don't share properties in common that are useful for scientific generalisation and prediction. In the face of nominal reference concepts are discarded for scientific purposes. If the concept of mental disorder or a concept of a particular kind of mental disorder turned out to have nominal reference then we should eliminate that concept from science.

Another way that reference could go would be **split reference** where the concept refers to more than one category. The most often cited example of split reference is how our concept **greenstone** turned out to refer to two different categories: jadeite and nephrite. While in this instance we eliminated the concept **greenstone** from science there are other cases where we retain the concept such as when biologists conclude that there are two species of Tuatara.

Another way that reference could go would be if there turned out to be **partial reference**. In partial reference our concept is found to refer to refer

to more instances than we had taken there to be. When we learned that whales were mammals, for example, then we had to revise our concept of mammals. This is why it is important not to get too caught up in conceptual analysis when one is interested in the nature of the world. Another way that partial reference could go would be if our concept referred to a category but also a collection of other instances that turned out not to share generalisable properties with instances of the category.

Wakefield criticises the DSM for being too liberal with the criteria so that many individuals are diagnosed as being mentally disordered when they aren't. He argues this on conceptual grounds because it follows from his 'harmful dysfunction' analysis of the concept of disorder rather than because of the lack of generalisable properties, however. For our concepts to be maximally scientifically fruitful it would be best if they were revised so as to allow us to identify members of categories that share properties in common that allow us to make generalisations and predictions. In the face of partial reference we could eliminate our concept though it would seem more fitting to revise our concept so it falls in line with a category if there is one in the near vicinity.

We can thus see that if our concept of mental disorder turned out to have nominal, partial, or split reference then one could use this to motivate eliminativism. We also have concepts of particular kinds of mental disorder such as depression, obsessive- compulsive disorder, schizophrenia, autism, and the like. If one or more of these concepts turned out to have nominal, partial, or split reference then one could use this to motivate eliminativism about that particular kind or kinds of disorder. In the case of split reference scientists do sometimes distinguish between higher and lower categories and retain the concepts for the higher category. In the case of partial reference we would also not be forced to eliminate our concept, however, as we could instead revise our conception so that it did refer to a category.

Even if there is **full reference** where the concept fairly straightforwardly refers to a category there could still be grounds for eliminativism, however. In the rest of the seminar I want to consider the different kinds of categories that

could be relevant referents for our concepts of mental disorder and particular kinds of mental disorder and see which of these could lead us to eliminativism about our concepts.

Kinds of Categories: Part One

Essential Kinds are thought to be categories that share the same intrinsic, or non- relational essential properties. Paradigmatic examples include water and gold where in order to count as an instance of water the instance must have the property of being H2O and in order to count as an instance of gold the instance must have the property of being atomic number 79. The intrinsic properties are thought to be constitutive of kind membership. While essential kinds aren't particularly relevant for psychopathology it is easier to understand other categories by way of contrast.

A kind of category that would seem to have better prospects for psychopathology would be the category of **Biological Kinds**. Paradigmatic examples of biological kinds include elms and tigers. I shall consider two different accounts that have been offered of the nature of biological kinds before considering how biological kinds could be relevant to psychopathology. The first account is probably the most widely accepted and on this account the essential properties for biological kinds are thought to be relational, extrinsic properties of historical lines of descent.

Mallon considers a second account of biological kinds. He maintains that some theorists have regarded biological kinds as homeostatic property clusters. The notion here is that certain properties are found to be clustered together in nature. If we see some properties then we can infer the presence of other properties and thus homeostatic property clusters support scientific generalisations and predictions. Some theorists maintain that the reason why certain properties are found to be clustered together in nature is because they share some underlying causal mechanism that is responsible for the properties homeostasis. It is because the causal mechanism is found in the different instances that results in our being able to generalise and make

accurate predictions and if there were no common causal mechanism then we wouldn't have the generalisation and prediction power that we do.

This move would seem to lapse back to either essentialism or the relational historical view, however, depending on whether one attempted to cash out the relevant mechanism as intrinsic or extrinsic. It might be the case that intrinsic mechanism result in the tightest property cluster as in chemical kinds, for example, while extrinsic or relational causal mechanisms result in weaker property cluster as in biological kinds for example, though these property clusters are still useful for science. Woolfolk considers that some theorists maintain that biological kinds are homeostatic property clusters even though they do not share intrinsic or relational essential properties. These theorists aren't eliminativists about biological kinds, however, as they maintain that the biological property clusters are still useful for science with respect to generalisation and prediction.

There are a couple of interesting features of this latter view. The first is that it seems to be very much in line with the way the DSM carves up different kinds of mental disorder. I have already said how the DSM provides a number of behavioural symptoms and in order to meet criteria for a given disorder the person must meet some specified number of those symptoms. The majority of diagnoses do not have essential symptoms and thus members of diagnostic categories exhibit family resemblances of symptoms. This is similar to the property cluster view in that there can be some variation or family resemblance in the properties exhibited by individual members of the category. A feature of the property cluster view is that different instances have slightly different features and they may be more or less prototypical. Not all birds can fly, for example. There can also be borderline cases where it is unclear whether the instance is in fact an instance of the category or not.

Another feature of interest is that the properties that are of interest in the DSM are behavioural symptoms. One might thus consider the DSM to be treating mental disorders as **Behavioural Kinds** where each kind of disorder is a property cluster of behaviours. Theorists who adopt the property

cluster view as I have outlined it about biological kinds often take similarly superficial, observable properties to be the relevant properties while remaining agnostic as to the underlying causal mechanism. The category of birds, for example, includes such properties as flight and feathers where these properties are superficial properties rather than properties at a lower level of analysis such as genetic.

A problem that one might have with the homeostatic property cluster view as I have outlined it is that causal mechanisms seem to matter. We would like to know *why* it is that these properties are found clustered together and what caused the features or symptoms to maintain homeostasis. While an important part of science involves observation and description, as science progresses it starts to develop theories of the causal mechanisms responsible for the phenomena. Often our taxonomy is revised as we delineate kinds on the basis of causal mechanisms rather than superficial similarities. This is because if we classify on the basis of causal mechanisms the properties are likely to be better suited to generalisations and predictions.

Wakefield, in particular, pushes the intuition that causation matters and that even if there is a cluster of behavioural features if the 'right kind' of causes are absent then we do not regard the person as mentally disordered. One example he offers is a case of a person meeting DSM criteria for reading disorder. He maintains that if we find the person can't read because nobody ever instructed him how to read that we wouldn't regard the person as mentally disordered. If we found that the person had received adequate instruction and yet could not read because of some kind of inner malfunction, however, then we would conclude that the person was mentally disordered.

Wakefield thus characterises the 'right kind' of causes to be ones that are internal to the person. He is especially focused on the notion of neurological and / or cognitive malfunction which he characterises along the lines of a hardware / software distinction and while he doesn't mention it I don't think he would be opposed to adding genetic malfunction to the mix (supposing that it makes sense to talk of genetic malfunction). **Neurological kinds** would seem to be fairly straightforwardly thought of as biological kinds.

Some theorists have attempted to analyse **Psychological kinds** as another variety of biological kinds where mental or cognitive states such as belief and desire are the kind of state they are in virtue of what the mechanisms that support the state have evolved to do. Sometimes theorists (like Wakefield) appeal to current functions instead of evolutionary functions where the effects of a current function are responsible for the mechanism being prevalent in current populations. Treating mental kinds as biological kinds is controversial, however. In many respects they have more to do with some of the other varieties of kinds I want to consider: Socially constructed kinds.

Kinds of Categories: Part Two

Artefacts like pens and chairs are paradigmatic examples of **Socially Constructed Kinds**. Instances of the category pens could be thought of as instances of the category in virtue of having the relational property of being designed by an agent for a certain function. As such agents designing them for a certain function is a necessary and sufficient cause for category membership. Alternatively, one could characterise token pens as being instances of the category in virtue of having a cluster of properties in common. We certainly identify pens on the basis of these properties. The cluster of properties that the instances have in common in virtue of their being designed by agents allow for generalisations and predictions to be made about pens, though it might be that there are less of these available to us than there are about chemical kinds or biological kinds.

Socially constructed categories are distinctive from those other categories in the sense that people's intentional states are a necessary cause of the instances. Once the instances have been brought into being, however, then it is a mind independent fact that the instances are in fact members of the category. Even if our social practices changed so that we no longer used pens to write with or even if we lost our concept of a pen so that we couldn't identify instances as pens the instances of the category would continue to be instances of the category in virtue of their being designed by an agent with

a certain intention. As such the intentions of agents play a necessary causal rather than a constituent role with respect to category membership.

While pens and chairs are socially constructed categories it is hard to see why someone would want to be an eliminativist about our concept of a pen. While someone might want to eliminate pens in the sense of eliminating instances of the category and making people write with pencils or crayola crayons this is not a case of eliminativism about our concept PEN. If we find that a category is socially constructed we are not thereby required to be eliminativists. It is useful to have this notion of a socially constructed category as a backdrop for understanding some of the other socially constructed categories that might seem more relevant for mental disorder.

The notion of a **Looping Kind** was introduced by Hacking and this notion has subsequently been picked up on by other authors such as Griffiths and Mallon. In order to describe the features of looping kinds I need to draw a further distinction between what I shall call explicit looping kinds and implicit looping kinds.

Explicit looping kinds are kinds that are dependent on our social practices in the sense that the instances wouldn't have existed as instances of the category if our social practices had been different in certain respects from what they were. They are thus constituted by our social practices and they are different from artefacts in the sense that if we altered our social practices in certain ways then the instances would no longer share the properties that are characteristic of their category membership. It is easiest to see this by way of example. Members of Parliament and Licensed Dog Owners are examples of explicit looping kinds. The category Members of Parliament relies on our social practices not only with respect to the instances sharing properties in common but also with respect to the instances continuing to share properties in common. The category is constituted by our social practices. If we altered our social practices so that we no longer had parliament, for example, then while the instances of the category would continue to exist they would lose the properties that are relevant for membership in that category.

These looping kinds are explicit in the sense that we are aware that the categories are dependent on our social practices. We know that there wouldn't be any Members of Parliament if we altered our social practices in certain ways. While one might well want to eliminate the instances as instances of the category Member of Parliament by blowing up either parliament or politicians, for example, this is not eliminativism about our concept of Member of Parliament. People could similarly want to eliminate mental illness by curing it or by eugenic policies but this is also not the relevant notion of eliminativism. The relevant notion of eliminativism would be to advocate that we eliminate our concept because Members of Parliament are dependent on the continuation of our social practices and that causal mechanism is the wrong kind of causal mechanism for categories. We accept this same causal mechanism in the case of pens, however, and Members of Parliament share properties in common that are useful for the special science of politics and thus we are not required to be eliminativists about explicit looping kinds. Indeed, eliminativism about members of parliament would seem to undermine politics as a special science.

Implicit looping kinds are similar to explicit looping kinds except in this instance we are unaware of their status as looping kinds and if we were to became aware of this then Hacking maintains our social practices would change and as a result the instances would no longer share the properties that support generalisation and prediction. Examples of implicit looping kinds include categories such as demonic possession and being possessed by a wild pig. The notion is that when our social practices legitimated these categories people came to behave in such ways and thus we could have pointed to the properties that instances of the category shared. Members of the category were identified as instances of the category because they shared certain properties in common. What is supposed to be distinctive about these categories, however, is that they cannot survive our realisation of their status as looping kinds.

The notion is that once we realise that these individuals display their common features in virtue of our social practices then we inevitably alter our

social practices so that the individuals no longer display those common features. This phenomena is probably best conveyed by way of Ian Hacking's characterisation of Multiple Personality Disorder which he takes to be an 'all too perfect illustration of the feedback effect' in implicit looping kinds:

> We tend to behave in ways that are expected of us, especially by authority figures – doctors, for example. Some physicians had multiples among their patients in the 1840's, but their picture of the disorder was very different from the one that is common in the 1990's. The doctors' vision was different because the patients were different; but the patients were different because the doctors' expectations were different. That is an example of a very general phenomenon: the looping effect of human kinds. People classified in a certain way tend to conform to or grow into the ways that they are described; but they also evolve in their own ways, so that the classifications and descriptions have to be constantly revised. (Hacking, 1995, p. 21).

Hacking maintains that in the case of implicit looping kinds there is a tension in that our social practises are the mechanism that both stabilises and destabilises the property cluster. With respect to the stabilising function he considers that individuals symptoms are shaped because when the clinician applies the concept to the patient this results in the clinician having either implicit or explicit expectations of the symptoms they expect to find in the patient. This changes the way that the clinician relates to the patient and is thought to lead to the patient exhibiting the symptoms they are expected to exhibit. Another way this can happen is if the clients apply the concept to themselves and thus come to exhibit symptoms that are stereotypic features of the concept. In this way the concept and our social practices stabilise the symptoms that the patient exhibits as they come to behave in ways that are consistent with the stereotype.

This story seems to be causal, but it is also thought to be constitutive in the sense that Hacking maintains that if we become aware that a category is a looping category then this will lead to our changing our expectations

and social practices and thus the properties will no longer occur. While our intentions were thought to be necessary causes of artefacts our intentions could alter and artefacts would continue exhibit the properties in virtue of which they are members of the category. Thus, while our intentions are necessary causes of the properties that artefacts have they are not sustaining causes of those properties. Our intentions can alter but pens continue to display the properties in virtue of which we can make generalisations and predictions about them. With looping kinds our concept seems to play a causal role once more but in this case it may be thought to be constitutive in the sense that our intentions and social practices are sustaining causes of the properties the instances have in common. If our intentions and social practices alter then the instances no longer exhibit the properties in virtue of which the instances were members of the category.

Hacking thus also considers how our social practices can have a destabilising effect. He traces how the stereotypical features of Multiple Personality Disorder have evolved through time. Hacking tells a complex story of destabilisation and he draws on a variety of factors including political and theoretical, which lead to the concept evolving and the symptoms evolving in response to the evolution of the concept. Some examples he has of this effect in the case of MPD include how many alters are thought to be typical (one or several or over one hundred); whether there is one or two way amnesia; how long it takes to switch between alters; and reports of abuse. It thus seems that the change seems mostly to be a function of a change in the theoretical views of clinician's. This led to a subsequent change in how they related to their clients and what kinds of symptoms they were interested in seeing. Hacking seems to regard implicit looping kinds as having some homeostasis but the homeostasis is less stable than other kinds of socially constructed and natural kinds perhaps because our concepts evolve much faster.

In these cases because it is implicit that we are dealing with a looping kind we are unaware of the impact of categorisation, our social practices, or our expectations, our ways of interacting with the person, and so forth. If we come to believe that a certain kind of mental disorder is a looping kind then

it seems that one of three things could happen:

Firstly, it could turn out to be the case as an empirical matter of fact our change in belief does not result in a change in our social practices. While Hacking thinks the relevant social practices are ones that invariably would change if we became aware that the category was a looping kind surely it could be possible that the social practices that are sustaining the phenomena could be resistant to change possibly because they have other beneficial effects. It is unclear whether Hacking would consider this to be an example of an implicit looping kind because it was implicit even though awareness did not result in it dissolution or whether Hacking would consider this to be an example of an explicit looping kind because it does not dissolve in the face of our awareness even though the so called explicit looping kind was implicit for a time.

Secondly, it could turn out to be the case that as an empirical matter of fact that if we came to believe the category was looping and we changed the relevant social practices the properties remain. In this case we seem to be left having to conclude that the category wasn't a looping kind after all. While it could still be socially constructed in the sense that artefacts similarly rely on us for their initial existence the phenomenon wouldn't seem to be maintained by our social practices.

The third thing that could happen would be that the defining properties of the category could shift so that there wouldn't be any properties that the instances shared that were of any use for scientific generalisation or prediction. In this latter case we would be left with a **Nominal Kind**. Nominal kinds aren't really kinds at all from a scientific point of view and thus we would eliminate the concept from science. It thus seems that we will end up being eliminativists about implicit looping kinds if implicit looping kinds are kinds such that being aware of their status is enough to effect social change which is enough to destabilise the property cluster so that it is no longer scientifically fruitful.

This is the way Hacking characterises them though it seems that those di-

mensions might be teased apart. If we found that a particular kind of mental disorder was an implicit looping kind this isn't to say that all instances of the category are suddenly cured of all symptoms of psychopathology, however, it is just to say that they won't display features of psychopathology that were stereotypic of the looping kind. Eliminativists about Multiple Personality Disorder often say that there is no such thing as Multiple Personality Disorder there is only Borderline Personality Disorder that has been worked up into Multiple Personality Disorder in response to our social practices around our concept of Multiple Personality Disorder. The notion here seems to be that if we refuse to participate in those social practices the patients will display stereotypic features of Borderline Personality Disorder. What is unclear, however, is whether this would be so because the clinician's expect them to come to display the stereotypical features of Borderline Personality Disorder or whether this is in response to some other mechanism.

While Multiple Personality Disorder is one of the favourite categories of those who maintain we should be eliminativists it is unclear whether other, more paradigmatically biological psychiatric kinds could turn out to be looping kinds or to have a looping kind feature to the behavioural symptoms. I want to end with a question: If looping kinds are the kinds that mental disorders often are, then what are the consequences for the science of psychiatry?