Harmful Dysfunction and Mental Illness: Why the latter is not the former

Leilan Nishi*
lnishi@scu.edu

Abstract

In his essay 'The Concept of Mental Disorder' (1992), Jerome C. Wakefield puts forth a hybrid account of mental disorder that relies on the concept of 'harmful dysfunction', wherein 'harmful' is a subjective value term determined by social norms, and 'dysfunction' is the objective value-neutral counterpart that denotes the failure of a mechanism to perform as evolutionarily intended. In this paper, I begin by laying out the kind of commitments Wakefield is wedded to, which will demonstrate that Wakefield's 'harmful dysfunction' account of illness critically fails to unify accounts of physical and mental illness. I claim this because the concept of mental dysfunction itself is not value-neutral like Wakefield assumes and needs it to be, which makes the view unworkable when applied to mental disorder. In its place I propose a model of human flourishing that will account for many different models of mental functioning.

1 Introduction

Jerome C. Wakefield (1992) undertakes a grand venture when he seeks to come up with a comprehensive account of disorder that can explain what physical and mental disorders have in common. He defines disorders as 'harmful dysfunctions': 'harmful' being a value term determined by social norms, and 'dysfunction' a scientific term referring to the failure of an evolutionarily designed mental mechanism to function as intended (1992, 373). In certain ways, he succeeds. By combining the apparent opposites of value-laden and scientific approaches, he avoids the issues that come with defining disorders as either entirely normative or entirely descriptive.

However, his concept of mental disorder as harmful dysfunction contains a fatal problem: it cannot be adequately applied to mental disorders. In this paper, I will argue that there is reason to doubt, and ultimately reject, aetiological dysfunction as a marker of disorder, because

*Leilan Nishi is currently finishing the last year of her undergraduate degree in Philosophy at Santa Clara University. While her main interests are in philosophy of mental illness and existentialism, she has yet to encounter a philosophical field she doesn't like. She will be starting her PhD at CUNY Graduate Center in the Autumn.

it is in fact not value-neutral in the way that Wakefield assumes. I will argue that Wakefield's appeal to aetiological function and rational agency contains commitments to two implicit norms: biological determinism and a normatively loaded conception of rationality. I end by gesturing at an alternative model of function, which accounts for the multitude of varying ways in which people function and promotes a life of flourishing.

2 The harmful dysfunction model

In this section, I will briefly lay out the commitments embedded in Wakefield's 'harmful dysfunction' account of disorder. The harmful dysfunction (HD) analysis is a hybrid account of disorder. The dysfunction component is the scientific half of the formula and is grounded on the concept of etiological (i.e., evolved) function: a mechanism is dysfunctional when its present state of operation has deviated from its intended function as determined by evolution (1992, 374). The harmful component is a value-laden concept that refers 'to the consequences that occur to the person because of the dysfunction and are deemed negative by sociocultural standards' (Wakefield 1992, 374). Only those mechanism malfunctions that are socially disvalued can be considered disorders. The HD analysis provides a concept of disorder that is meant to apply to all conditions we call disordered, thereby unifying them under a single category definition. The result is intended to be a cohesive account of disorder that explains why certain conditions are not merely environmental issues or socially disvalued conditions.

Wakefield defines a function more broadly as an effect that explains its cause: the heart's effect of pumping blood, for example, enters into the explanation of its cause for existing in our bodies, which makes pumping blood the function of the heart (he also uses the examples of seeing as the function of the eyes, and mobility as the function of legs, using this same reasoning). Aetiological explanations for natural functions provide a causal story for how the mechanisms possessed by an organism that contributed to its reproductive success become naturally selected and present in us today. He wants to extend such etiological explanations for physical mechanisms like pumping blood, seeing, and walking to *mental* mechanisms:

> Considering that mental processes play important species-typical roles in human survival and reproduction, there is no reason to doubt that mental processes were naturally selected and have natural functions, as Darwin himself often emphasized [(Boorse 1976)]. Because of our evolutionary heritage, we possess physical mechanisms such as livers and hearts; that same heritage gave us mental mechanisms such as various cognitive, motivational, affective, personological, hedonic, linguistic, and behavioral dispositions and structures. Some mental conditions interfere with the ability of these mental mechanisms to perform the functions that they were designed to perform. In such cases, there is a part dysfunction of the particular mental mechanism. (Wakefield 1992, 375)

Wakefield assumes, without justification, that there must be an etiological story that explains

the existence of mental mechanisms like there are for physical ones; he seems to imply that because evolutionary heritage explains the physical, it is necessary that it explains the mental, because, he argues, it couldn't be merely a happy accident that we developed structures and dispositions that work so intricately and harmoniously together so well to provide the ultimate remarkable benefit of rationality. I argue that, even if we grant him this claim[1], his account of mental dysfunction is unworkable due to the normativity inherent in what is supposed to be a descriptive account of mental mechanism aetiology.

3 The problem of rational agency

Wakefield's account of the etiological function of mental mechanisms relies on his own unjustified conception of rational agency as value-free. When Wakefield discusses dysfunctional mental mechanisms, he almost universally appeals to his own conjecture that mental mechanisms were evolved to make us rational. Consider the following two examples of dyslexia and depression.

Wakefield (2000) uses dyslexia as a supposedly obvious example of a dysfunctional mental mechanism. He says, in regards to individuals who cannot learn how to read, that with those who 'seem incapable of learning to read even under optimal learning conditions, we infer that there is something wrong with some internal neurological mechanism that, when functioning as designed, supports the capacity to read (although it supports reading accidentally, not by design)'[2] (2000, 256). Again, his statement is rife with assumptions about what this 'some neurological mechanism' is that is malfunctioning, which betrays his own lack of clarity on whether or not there even is a mechanism, and if there is, what it is. More notably, there is another assumption at work: that there exists an evolutionary dysfunction in this instance at all. He simply takes for granted that in the case of an individual who couldn't learn to read under conditions conducive to learning how to read, we would automatically infer that there was 'some internal neurological mechanism' that is failing to function as designed. The implication is that we evolved with the capacity to read, which is 'an invented way of exploiting our selected mechanisms for our own purposes' (2000, 255), the purpose being rationality. For Wakefield,

1. In addition, to move forward we must also ignore two extremely damaging epistemic concerns resulting from his etiological commitment that specific mental mechanism have specific evolved functions: (1) how do we identify the mechanism we should be looking at when a potential mental malfunction occurs for a specific condition? (2) even if we could solve that problem, how do we identify the evolutionary function of these mental mechanisms? Wakefield seems to realise that these are issues (2000, 263–64), but provides no way of solving them.

2. Wakefield here is talking about spandrels, or un-designed side effects of design, which he argues can malfunction in a harmful way when they are inevitable by-products of design, in which case the failure of the spandrel implies the failure of some intended function. He does not explain how this could apply to mental illnesses, but the following example would be in line with his theory: suppose there is an evolved mechanism for symbol recognition, and the inevitable spandrel of that design is the capacity to read written language. He would argue that a failure of the spandrel of reading is indicative of a failure of the evolved mechanism of symbol recognition. However, he would still be faced with the two epistemic issues mentioned in the first footnote, which remain just as practically insurmountable.

not being able to read is not merely a difference, but a dysfunction, because illiteracy marks an obstacle to exploiting our selected mechanisms towards this purpose of making us better able to reason and interact with our world to logically pursue our ends.

In Wakefield's particularly damning treatment of depression (2000, 266), we find another example of his belief that rationality is an evolutionarily selected function. He calls depression disordered when the 'loss [response is] extremely disproportionate to experienced loss,' which, as he acknowledges, assumes that 'sadness as a designed response to loss [could] turn out to be incorrect.' He explains away cultural values and culturally defined expectations for expressing sadness by arguing that depressive behaviour disvalued in one place and valued in another could have normal and abnormal sources; in other words, while 'there may be experiences when growing up (or, for all we know, genetic dispositions) that cause people in those cultures to express sadness more readily or intensely or more enduringly than in our culture, [someone] in our culture who has not had those experiences and yet reacts with a similarly high level of intensity may be doing so for entirely different reasons, including possibly a dysfunction' (2000, 266). Furthermore, he says that we should adjust our judgements and attributions of disorder according to other circumstances regarding the loss that could explain why the individual is reacting so strongly: 'if the individual's personality, the special meaning of the loss, or other circumstances suggest that a more intense or enduring response than the usual is due to non-dysfunction factors, we refrain from attributing disorder.'

However, Wakefield does not explain where those explanations end; he gives no guide for how we can differentiate between a personality that is inclined to dramatic expressions of emotion and a disorder, or how we know if we have correctly assessed the correct significance of the loss to the individual. He merely says that 'we try to judge when the reaction goes so far out of the usual bounds that it seems unrelated to any possible coping benefits', in which case '[w]e then become more persuaded that there is a possible dysfunction' (2000, 266). He assumes that if there is no identifiable culturally learned explanation for why someone is responding with extreme depression to a certain loss, we would not be unreasonable to infer a dysfunction. Similarly, he assumes that in lieu of personal circumstances related to the loss that could contribute to a 'disproportionate' loss response, we could potentially infer a dysfunction as well. He is assuming that, if there are no external or personalised circumstances that could explain the extreme sadness, we can infer a dysfunction. There is a notable and damaging implication from these assumptions: both of these limitations on attributing disorder are limits to determine when 'the reaction goes so far out of the usual bounds that it seems unrelated to any possible coping benefits', which would indicate that 'the disorder can be recognised by the fact that there is great sadness for no apparent reason'. He concludes by saying that 'none of this complex, contextually anchored reasoning, however speculative and fallible it may be, has anything inherently to do with local values' (2000, 266).

The problem is that it does in fact have to do with values, and that is the value of rational agency that underlies his assessment of how and when we recognise depressive disorder. When he refers to sadness that is 'disproportionate' and outside of 'usual bounds' as disordered, he cannot be saying they are statistically deviant modifiers, as those would not be aetiological cate-

gories. Wakefield is referring to rational agency. The sadness response is disproportionate in the sense that is unreasonable: the individual is failing to assess the magnitude of the loss correctly, and so reacts in a way that is unresponsive to the facts, much like the case of the chronically low and unresponsive self-esteem. The rational response would be to change one's sadness to match the magnitude of the loss, so when that does not happen, there is a failure to be rational-and for Wakefield, a disorder. In other words, feeling overwhelming sadness is a disorder when the magnitude of the sadness is beyond the rational range of response in regards to the loss, thereby constituting a reduction in the ability to be rational. We can make sense of Wakefield's view on dyslexia, depression, and self-esteem only through this notion of mental dysfunctions as undermining rational agency.

Rationality as the metric would not be an issue if it could reliably differentiate disorder from non-disorder while being value-neutral, as Wakefield believes. However, neither of these demands are met by the kind of rational agency that Wakefield assumes. The conditions that are disorders are those that are indicative of an impingement on rational agency—specifically, a condition is an impingement when it '[tends] to cause a person to act contrary to their interests without an adequate reason for doing so, and [impairs] a person's ability to decide competently and voluntarily, for example, by disrupting one's cognitive abilities' (Edwards 2009, 79). If we ask how we differentiate between a character predisposition towards 'intermittent but massive and harmful lapses in rationality' (2009, 80) and a mental illness, we find that the distinction is made on normative judgements, not descriptive ones. This distinction cannot be attributed to statistical distribution in a population, environmental circumstance, or severity (2009, 80).

In this same way, rationality is normative. There is no blueprint for what rationality should look like, both in mode of function and degree of function. To champion one model of rationality over another is to make a normative judgement; the parameters and applications of this rationality are arbitrary. Consider two of Edwards' examples: first, the description of 'a couple living in a war zone during a bombing raid, frozen in fear under a precarious cover just a few meters from a bomb shelter that would greatly increase their chances of survival' (2009, 80), who, by not running to the bomb shelter, are failing to think rationally; second, of a youth who engages in behaviour that could not be described as rational, such as committing robberies where the 'potential takings could not possibly justify the risk' or committing assaults where there is 'no chance of avoiding arrest' (2009, 80). Edwards calls the bomb raid couple irrational, but not disordered, and the youth irrational, but morally responsible for their actions. This is because there are many mental conditions that impair our rational agency and can have a negative impact to our wellbeing, but are not considered mental illnesses, and because the way we differentiate between a character predisposition towards 'intermittent but massive and harmful lapses in rationality' (2009, 80) and a mental illness is based on morality. Now contrast that with Wakefield's self-esteem example: he infers a dysfunction and attributes disorder when a person has chronically low self-esteem that is not aligned with the facts about themselves (facts that should make them feel otherwise) and is unresponsive to those facts—in other words, when it is irrational. Why is his version of rational agency any more objective than the kind being impaired in these two examples? There is no framework to which he is appealing that can explain

why certain lapses in rationality, even serious ones, are indicative of a dysfunction, while others are understandable, acceptable, or a matter of character. At best, he is making a normative value judgement about the 'correct' understanding of the concept of rational agency, and at worst, an appeal to his own personal intuition.

Physical illnesses have none of these problems, because physical illnesses are not identified by their effects on rational agency. As a result, the HD analysis applies without trouble to examples of physical illness, which Wakefield realises, as he uses many physical examples to better explain his position. However, what he seems not to fully realise is that the extension to the mental realm cannot be made, as the examples he gives to defend his model for mental illnesses are overwhelmingly examples of physical illness: human chin and jaw (2009, 255), appendicitis (2009, 256), fever and morning sickness (2009, 259, 262), and sickle cell anaemia (2009, 260), *et al*. This means that the extension of the harmful dysfunction concept from physical illnesses to mental illnesses is utterly broken, and therefore it cannot apply to mental illnesses.

4 Why aetiology for mental mechanisms?

Suppose now that somehow it were possible to solve all these issues and maintain the harmful dysfunction model for mental illnesses. While impressive, there still looms the question we should have asked first: why should we use aetiology to determine the function of mental mechanisms? As previously stated, aetiology does not pose practically intractable problems for physical illnesses. This is because we can differentiate between difference and dysfunction by referring to a basic account of biological determinism that is not value-laden. This is not functional determinism, which is the idea 'that functions take place in a uniform mode at a relatively uniform performance level by a statistically distinctive portion of the members of a species' (Amundson 2000, 36)—this is problematic because considering functional mode-how an organ functions-fails to account for the fact that a comparable performance level can be achieved through a different mode of function. I propose instead a model of biological determinism to understand physical illness: the idea that certain organs evolved with respect to other organs within a biological system in order to create an optimally functioning organism, and that when the organ does not develop as intended, there is a biological dysfunction, which does not necessarily have to result in reduced performance level or ability. By adhering to aetiology for physical conditions, we get the benefit of the unification of conditions under the harmful dysfunction model as a complete and comprehensive way of understanding and classifying conditions that stays faithful to the value-laden and value-neutral formula.

There is no such benefit conferred from applying an aetiological account of mental function. This is because we are persons, and 'the interesting thing about *persons*, and possibly other things that have sophisticated mental lives, is that we value things other than survival and reproduction, and for the most part we evaluatively judge that other people *should* value things other than survival and reproduction (e.g., happiness or fulfillment)' (Edwards 2009, 77). Aetiological

accounts of function not only fail to capture what it is we care about; they stifle such understandings. I have already shown this with rationality, but the same is true of a misapplication of biological determinism to the mental realm. This is because there are no blueprints for mental mechanisms like there are for physical mechanisms; there is no objective standard for what makes a self-esteem mechanism (assuming there is one) that is low more or less dysfunctional than one that is high. Even that mode of function that is non-typical 'is not broken by its failure to comply with some imagined blueprint [...] It will function anyhow, in spite of its atypicality', because there is no blueprint for mental function. Even if there was a blueprint, that blueprint would only be one that brings about functional integration, so that various mechanisms develop together and adjust to one another in order to function, as evidenced by the incredible multitude of ways in which people function (Amundson 2000, 39). Functional and biological determinist accounts have the same problem as models of rationality: though touted as descriptive, what is really happening is a normative judgement about the desirability of certain modes of mental mechanism function over others.

The only type of function we can employ to adequately encompass all these different kinds of mental functioning without sacrificing variation is a Cummins function (Woolfolk 1999 665–67). A Cummins or ahistorical function is one that focuses on 'the causal relations among systems and their component parts, such that "the function of a part of a system is its causal contribution to some specified activity of the system" [(Walsh and Ariew 1996, 493)]'. This means that the Cummins function of a certain mechanism is defined in relation to a particular designated purpose of the system as a whole, which could be entirely arbitrary; it is 'interest relative,' and so 'many different systems can be posited that concurrently contain the component' (1999, 665). Because the designation of the system is arbitrary, 'no background context of inquiry is privileged over any other, as is the case with the privileging of an evolutionary account by historical functional analysis' (1999, 666). A mechanism is functioning if it is concurrent with whatever framework of interest is being applied, and failing to function if it is not; this means that the same mechanism can also be dysfunctional according to one account while being functional in another. For example, to take Woolfolk's example of the heart (1999, 665–66), in the context of explaining human physiology, the Cummins function of the heart is to pump blood; in the context of an electrocardiogram, to produce electrical signals that result in EKG tracings; in the context of assassination, to bleed and lead to death.

For mental mechanisms, we can privilege certain background contexts, because the one we privilege is going to be that which best explains the interest we have. This means understanding their functionality in different contexts according to certain frameworks of interest—and these frameworks are not going to be aetiological ones. Aetiological frameworks could give us explanations only if our interest is in the biological background and history of mental mechanisms (granting for the moment that it is even possible to determine); such concerns are in the realm of theory, and so remain practically removed from our current environment, giving us no understanding of how to value these mechanisms in a practical manner. In Wakefield's own words, '[t]he mental health theoretician is interested in the functions that people care about and need within the current social environment, not those that are interesting merely on evolutionary

theoretical grounds' (1992, 384). Human flourishing is the interest.

5 Conclusion

Wakefield's harmful dysfunction analysis, while an excellent model for classifying physical disorders, cannot be applied to mental disorders as he believes. The HD analysis of mental disorder cannot escape from the fact that the idea of a mental mechanism dysfunction is grounded in the normative concept of rationality, as are the concepts of aetiology and biological determinism. We therefore must abandon such historical concepts of mental function and adopt a model that can account for mental variability. This model will be an ahistorical account of interest-relative function that accurately captures the interest that really matters to us: how to live our best and most meaningful life. Because this will look different for each person depending on who they are, this new framework will allow for any number of different lifestyles, none of which are dysfunctional merely because they are different. It will make room for these alternative modes of being while understanding that mental conditions, including serious ones, are not pathological but struggles to be overcome on the path to leading the most fulfilling life.

References

Amundson, Ron. 2000. "Against Normal Function." *Studies in History and Philosophy of Science Part C* 33 (1): 33–53. doi:10.1016/s1369-8486(99)00033-3.

Boorse, Christopher. 1976. "What a Theory of Mental Health Should Be." *Journal for The Theory of Social Behaviour* 6 (1): 61–84. doi:10.1111/j.1468-5914.1976.tb00359.x.

Edwards, Craig. 2009. "Ethical Decisions in the Classification of Mental Conditions as Mental Illness." *Philosophy, Psychiatry, & Psychology* 16 (1): 73–90. doi:10.1353/ppp.0.0219.

Wakefield, Jerome C. 1992. "The Concept of Mental Disorder: On the Boundary Between Biological Facts and Social Values." *American Psychologist* 47 (3): 337–88. doi:10.1037/0003-066x.47.3.373.

———. 2000. "Spandrels, Vestigial Organs, and Such: Reply to Murphy and Woolfolk's "The Harmful Dysfunction Analysis of Mental Disorder"." *Philosophy, Psychiatry, & Psychology* 7 (4): 253–69.

Walsh, Denis M., and André Ariew. 1996. "A Taxonomy of Functions." *Canadian Journal of Philosophy* 26 (4): 493–514. doi:10.1080/00455091.1996.10717464.

Woolfolk, Robert L. 1999. "Malfunction and Mental Illness." *Monist* 82 (4): 658–70. doi:10.5840/monist199982429.