Opinion-Perspective Article

A Harm-Based Framework for Defining Mental Illness: Moving Beyond Wakefield's Dysfunction Model

Charles Page

Charles William Page, 200 Seminary St, Pennsburg, PA 18073, United States

ABSTRACT

The definition of mental illness remains contested within the philosophy of psychiatry. Jerome Wakefield's "harmful dysfunction" account provides valuable insight by combining scientific and evaluative components. However, his reliance on biological dysfunction is problematic, as it excludes conditions without evolutionary dysfunctions and fails to separate scientific and value-laden components of mental illness. This article proposes a new framework for mental illness, defining it for functional rather than folk purposes. It argues that mental illness for psychiatric intervention should be defined as any mental framework that harms an individual or others more than psychiatric intervention would. This harm-based model removes biological dysfunction from the definition, focusing instead on the comparative evaluation of harm and benefit. It preserves Wakefield's emphasis on value judgments while offering greater practical applicability and inclusivity. By removing dependence on the "selected effect" view, this definition accounts for disorders without evolutionary dysfunction, such as dyslexia and aggression disorders derived from naturally selected processes. Ultimately, this framework provides a more functional basis for mental illness in society, emphasizing harm reduction and clinical utility over evolutionary essentialism.

Keywords: Mental Illness; Psychiatry; Wakefield; Philosophy of Science; Harm Reduction; Biological Dysfunction

INTRODUCTION

The DSM-5-TR classifies mental illness as "a syndrome characterized by clinically significant disturbance... that reflects a dysfunction in the psychological, biological, or developmental processes" (1). While this serves as the primary diagnostic indicator in clinical settings, its reliance on "dysfunction" is the subject of a long-standing debate between naturalists, who view illness as an objective biological failure (2), and normativists, who argue that illness is a social construct based on undesirable behavior (3). Jerome Wakefield's influential "harmful dysfunction" account attempts to refine this by grounding it in evolutionary biology (4). Although this account improves on the DSM definition by acknowledging the value-laden nature of "harm," it ultimately fails to withstand scrutiny regarding its biological criteria.

The central stance of this article is that the "dysfunction" requirement is an unnecessary theoretical burden that limits clinical utility. The objective of this

Corresponding author: Charles Page, E-mail: cpage@perkiomen.org.

Accepted December 29, 2025
https://doi.org/10.70251/HYJR2348.41101104

manuscript is to develop a new account of mental illness that addresses the issues in Wakefield's account while retaining the benefits it offers. It is proposed that mental illness for psychiatric intervention be defined as any mental framework that harms an individual or others more than psychiatric intervention would. Under this framework, the classification of "disorder" is contingent upon the comparative safety and efficacy of available clinical responses.

THE HARMFUL DYSFUNCTION MODEL AND ITS LIMITATIONS

To understand the necessity of a new framework, one must first analyze why the current dominant model fails. Wakefield defines mental disorder as a "harmful dysfunction," a concept composed of a factual component (dysfunction) and a value-laden component (harm) (4). He grounds "dysfunction" in the "selected effect" account of evolutionary biology, championed by philosophers like Larry Wright and Ruth Millikan (5, 6). In this view, a mechanism is functional only if it performs the specific task for which it was naturally selected in the ancestral environment. To illustrate this, the example of the heart is often used: the heart was selected to pump blood, not to make a rhythmic thumping sound (5). Therefore, a heart is only dysfunctional if it fails to pump blood; it is not dysfunctional if it fails to make a sound (5). Wakefield argues that mental mechanisms follow the same logic: a condition is a disorder only if there is an objective failure of an evolved mechanism.

Despite its theoretical elegance, this reliance on the "selected effect" view creates significant gaps in clinical applicability. The primary problem is that biological dysfunction is strictly required, which forces a conceptual wedge between clinical reality and evolutionary history. This requirement excludes conditions that are clinically relevant but do not represent a failure of evolution. For instance, dyslexia is widely treated as a disorder (7), yet reading is a cultural invention too recent to be a "selected effect" of evolution (8). Consequently, the neural mechanisms involved in reading cannot be "dysfunctional" in an evolutionary sense, as they were never selected for literacy (8). Similarly, traits like high aggression may have been functional for survival in the ancestral environment but are maladaptive in modern society (9). In these cases, the mechanism is performing its evolutionary function perfectly, yet the outcome is pathologically harmful in the contemporary context. Furthermore, as Murphy and Woolfolk argue, deciding whether a grief response is "disproportionate" relies on social norms about grief, not just on biological facts (10). Deciding what constitutes an appropriate level of grief to experience involves a value judgment before the notion of harm is even introduced, suggesting that Wakefield's attempt to separate objective "dysfunction" from subjective "harm" is conceptually flawed.

A HARM-BASED FRAMEWORK FOR PSYCHIATRIC INTERVENTION

Since determining a universally accurate "folk" concept of mental illness is fraught with philosophical difficulty, this paper proposes a pragmatic, stipulative definition explicitly designed for clinical contexts. It is suggested that mental illness for psychiatric intervention be defined as any recurring, predictable mental framework that harms an individual or others more than psychiatric intervention would.

"Harm," under this definition, refers to phenomenally felt suffering—the subjective experience of "what it is like" to undergo distress (11)—by an individual or direct safety threats to those around them. Crucially, this harm must outweigh the potential adverse effects of psychiatric treatment (side effects, stigma, loss of autonomy). This definition introduces a comparative metric: a condition is only an "illness" if the "cure" (intervention) is less harmful than the condition itself. This inherently prevents over-pathologizing; if a condition causes mild distress but the available medication has severe side effects, it does not meet the threshold for "illness requiring intervention."

To address the risk of pathologizing non-normative behaviors, this framework requires "phenomenally felt badness" or direct safety threats. This prevents the historical errors of psychiatry, such as the pathologization of homosexuality, which was classified as a disorder until 1973 despite a lack of inherent phenomenal distress or safety threats (12). Societal disapproval alone does not constitute "harm" under this model. Suppose a trait does not cause the individual internal suffering and does not threaten the physical safety of others. In that case, it falls outside the scope of psychiatric intervention, regardless of its statistical rarity or social unpopularity.

IMPLICATIONS FOR DIAGNOSIS AND NEURODIVERSITY

This framework offers several advantages over the Harmful Dysfunction analysis. First, it resolves the

"Evolutionary Gap." By removing the requirement for biological dysfunction, this definition inclusively accounts for disorders like dyslexia (7, 8) and maladaptive aggression (9). The focus shifts from "what was this designed to do?" to "what is this doing to the patient now?" Second, it aligns with the Neurodiversity Movement. Critics of the medical model often argue that conditions like Autism or ADHD are differences, not deficits (13). Under the proposed harm-based model, these traits only qualify as "illness" if they cause suffering that exceeds the cost of intervention. If a neurodivergent individual is functioning well and not distressed, the trait is a "mental framework," not an "illness," thereby respecting patient autonomy. Recent scholarship reinforces this view, arguing that what is typically labeled 'dysfunction' is often a friction between diverse cognitive styles and the rigid demands of a 'neuronormative' system, rather than an internal biological failure (14). Third, it serves as a "Check and Balance" on Psychiatry. By baking the "harm of intervention" into the definition, this model forces a constant re-evaluation of medical practices. If a treatment is harmful, fewer conditions qualify for help, shifting the focus from evolutionary essentialism to clinical utility.

CONCEPTUAL BOUNDARIES AND LIMITATIONS

While this framework solves the evolutionary puzzle, it faces practical challenges. Determining "harm" remains partially subjective; psychiatrists may disagree on whether a patient's distress warrants the risks of medication. Additionally, ego-syntonic conditions (like the manic phase of Bipolar Disorder) present a challenge, as the patient may not feel "bad" in the moment. In these cases, the definition relies on the "harm to others" clause or the predictable future harm to the individual's own safety and interests. Furthermore, while this framework suggests that "shyness" could become an illness if a safe "cure" existed, the requirement for "phenomenally felt badness" or significant distress prevents the pathologizing of normal variations that do not cause significant suffering.

CONCLUSION

Wakefield's harmful dysfunction analysis bridged the gap between biological facts and social values, but its reliance on evolutionary history rendered it inadequate for modern clinical reality. This paper has argued for a shift toward a functional, harm-based definition: a subject is mentally ill for intervention only when their mental framework causes harm that outweighs the costs of that intervention. This approach decouples psychiatry from the ancestral environment, allowing for the recognition of modern maladaptations like dyslexia while protecting neurodiverse populations from over-pathologization. By focusing on phenomenal suffering and practical outcomes, this framework provides a more ethical and clinically practical guide for psychiatric practice.

ACKNOWLEDGEMENTS

I would like to thank Professor Aiden Woodcock of the University of Cambridge for his guidance in developing this paper.

FUNDING SOURCES

The author declares that no funding was received for the preparation of this manuscript.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work.

REFERENCES

1. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed, text rev. Washington, DC: American Psychiatric Association; 2022. https://doi.org/10.1176/appi.books.9780890425787
2. Boorse C. Health as a theoretical concept. *Philos Sci*. 1977; 44 (4): 542-573. https://doi.org/10.1086/288768
3. Szasz TS. The myth of mental illness. *Am Psychol*. 1960; 15 (2): 113-118. https://doi.org/10.1037/h0046535
4. Wakefield JC. The concept of mental disorder: On the boundary between biological facts and social values. *Am Psychol*. 1992; 47 (3): 373-388. https://doi.org/10.1037/0003-066X.47.3.373
5. Wright L. Functions. *Philos Rev*. 1973; 82 (2): 139-168. https://doi.org/10.2307/2183766
6. Millikan RG. Language, thought, and other biological categories. Cambridge, MA: MIT Press; 1984. https://doi.org/10.7551/mitpress/4124.001.0001
7. Peterson RL, Pennington BF. Developmental dyslexia. *Lancet*. 2012; 379 (9830): 1997-2007. https://doi.org/10.1016/S0140-6736(12)60198-6
8. Thornton T. Mental illness. Cambridge: Cambridge

University Press; 2022.
9. Nesse RM. Good reasons for bad feelings: insights from the frontier of evolutionary psychiatry. New York: Penguin; 2019.
10. Murphy D, Woolfolk RL. The harmful dysfunction analysis of mental disorder. *Philos Psychiatr Psychol*. 2000; 7 (4): 241-252.
11. Nagel T. What is it like to be a bat? *Philos Rev*. 1974; 83 (4): 435-450. https://doi.org/10.2307/2183914
12. Drescher J. Out of DSM: depathologizing homosexuality. *Behav Sci*. 2015; 5 (4): 565-575. https://doi.org/10.3390/bs5040565
13. Walker N. Neuroqueer heresies: notes on the neurodiversity paradigm. Fort Worth: Autonomous Press; 2021.
14. Chapman R. Empire of normality: neurodiversity and capitalism. London: Pluto Press; 2023. https://doi.org/10.2307/jj.8501594