Forthcoming in *Philosophy, Psychiatry, & Psychology*

Function, Dysfunction, and the Concept of Mental Disorder

Jonathan Y. Tsou

Naturalistic accounts of mental disorder aim to identify an *objective basis* for attributions of mental disorder. This goal is important for demarcating genuine mental disorders (e.g., schizophrenia, bipolar disorder) from artificial or socially constructed disorders (e.g., Drapetomania, homosexuality). The articulation of a demarcation criterion provides a means for assuring that attributions of 'mental disorder' are not merely pathologizing different forms of social deviance (or normal behavior). The most influential naturalistic and hybrid definitions of mental disorder (Boorse, 1976; Wakefield, 1992) identify *biological dysfunction* as the objective (i.e., factual) basis of mental disorders: genuine mental disorders are (necessarily) caused by biological dysfunction. There is disagreement regarding how to conceptualize the proper (or 'normal') mental functions that are disrupted in mental disorders.

In a rich and provocative paper, Anne-Marie Gagné-Julien (2021) argues that the causal role account of function defended by Robert Cummins (1975, 1983) can provide an objective foundation for the concept of mental disorder. Compared to Boorse's and Wakefield's (allegedly) value-free accounts of function, Gagné-Julien's account is explicitly value-laden insofar as it acknowledges the values and interests involved in ascriptions of function. She argues that this normative account of function can provide an objective basis for judgments of dysfunction (and hence, mental disorder) if it satisfies the social rules of objectivity articulated by Anna Alexandrova (2017, 2018). Gagné-Julien's paper is a welcome contribution to the literature on defining mental disorder, especially in its rejection of the anachronistic ideal of objectivity as value-freedom and application of an alternative social account. However, her resulting account of 'objective function' fails to yield the concrete kind of criterion required to address the demarcation problem.

The *social objectivity* that Gagné-Julien suggests normative accounts of function can achieve fails to directly address the problem of demarcating genuine mental disorders. Theorists who define mental disorder in terms of biological dysfunction aim to provide a *clear and concrete criterion* that distinguishes functional from dysfunctional mental traits. Gagné-Julien's analysis fails to present such a criterion. In particular, her causal role account of function fails to specify a clear standard of 'proper function,' which is needed to distinguish functional from dysfunctional mental capacities. Gagné-Julien endorses the causal role (CR) account defended by Cummins (1975), which conceptualizes functions as causal contributions of a component part to a capacity of a larger system. *Taken in its generic form*, a problem with Cummins' CR account is that—without specifying the *higher goals* of a system—it cannot identify the 'proper' (or 'normal') functions of systems and generates numerous pseudo-functions (Millikan, 1989; Neander, 1991; Griffiths, 1993). This difficulty arises because CR functions are *interest-relative* insofar as scientists *choose* to analyze capacities that are relevant to their fields of study

(Amundson & Lauder, 1994). Since Gagné-Julien is interested in functions relevant *for mental health*, she needs to indicate *what the higher goals of proper ('healthy') mental functions are* (i.e., mental capacities *for what*?). As a point of contrast, Boorse (1977) endorses a CR account that assumes that—*relative to the interests of medicine*—survival and reproduction are the highest goals of human organisms (pp. 555-556).[1] Accordingly, the proper (or 'natural') function of a biological part is its statistically typical ('normal') causal contribution to an organism's (current) capacity *to survive or reproduce*. Regardless of whether this is the right account of function relevant for health, it offers a *concrete criterion* that can distinguish functional mental capacities (i.e., causal contributions to biological fitness) from dysfunctional mental capacities (i.e., internal states that disturb these functions). The social account of objectivity that Gagné-Julien presents as *constraining* her favored CR account does not yield a concrete criterion of this sort. Rather, Alexandrova's analysis suggests that an objective theory of mental capacities should satisfy certain social conditions that ensure that the value presuppositions of the theory are subjected to critical scrutiny.

The considerations above highlight the fact that Gagné-Julien's analysis and Boorse and Wakefield's analyses assume incommensurable ideals of objectivity, which are suitable for different philosophical projects. The naturalistic accounts defended by Boorse and Wakefield aim to be objective in the sense of providing a factual (value-free) criterion for identifying proper biological functions. For Boorse, a biological function is the (statistically) normal causal contribution of an internal part to an individual's capacity to survive and reproduce. Wakefield endorses a narrower selected effect account, wherein a biological function is the beneficial effect of a mental mechanism that was naturally selected. Both accounts present a concrete naturalistic standard that permits objective (i.e., intersubjective) attributions of function. By contrast, the social ideal of objectivity endorsed by Gagné-Julien provides assessments on when the *practices of a scientific community* are objective.[2] If a theory of mental capacities satisfies Alexandrova's social rules, a scientific community is socially objective in the sense of *achieving a critically arrived at consensus* regarding which mental capacities (and corresponding CR functions) are relevant to mental health. While Gagné-Julien (2021) presents Alexandrova's account as a *method for constraining which mental capacities have proper functions* (§ 4), it offers no specific guidance on this matter. Rather, it prescribes the (prior) social conditions that an objective theory should satisfy, viz., value presuppositions are made explicit, these presuppositions are critically scrutinized for controversy, and relevant parties are consulted to settle controversies. Hence, Alexandrova's procedure specifies prerequisites for a socially objective theory, but it cannot constrain mental capacities in the manner needed (i.e., providing a concrete criterion for ascribing proper CR functions) to address the demarcation problem.

[1] In personal communication, Boorse indicates that his account of function is equivalent to a Cummins-style causal-role account, wherein the system outputs (or goals) of interest are survival or reproduction (cf. Amundson & Lauder, 1994; Boorse, 2002).

[2] Alexandrova's social account of objectivity follows a tradition initiated by Longino (1990). Whereas as previous accounts (e.g., Kuhn, 1977) equated objectivity with a *shared (intersubjective) set of epistemic values* (e.g., empirical accuracy, explanatory power) for choosing among competing scientific theories, Longino equates objectivity with the presence of a plurality of alternative scientific theories that promote *transformative criticism* (i.e., criticism of the implicit background assumptions and contextual values of competing theories).

As a more substantive problem, there is a worry that theories that survive the critical scrutiny of Alexandrova's social rules might merely reflect *prominent and entrenched prejudices or biases characteristic of some historical context*. This issue will be particularly salient for a normative theory of function, such as Gagné-Julien's proposed account. Just as values of the past supported dubious consensus views (e.g., homosexuality is a dysfunctional condition), values that are widespread in current societies (e.g., efficiency, productivity) might support equally questionable consensus views (e.g., ADHD is a dysfunctional condition). If certain values are *deeply entrenched in a historical context*, normative accounts of function may be unable to effectively prevent the *pathologizing of social difference*, which is the very issue that motivates the demarcation problem in the first place.

As a more general issue, it is unclear what exactly is at stake in Gagné-Julien's insistence that—in a definition of mental disorder—values should be located at the level *dysfunction* rather than *harm*. While Boorse argues that the *theoretical* concept of disease is value-free, he acknowledges that *practical* concepts of disease (e.g., employed in treatment contexts) are value-laden. In earlier works, Boorse (1976) argued that 'diseases' are factual (i.e., interferences to internal parts that play a standard causal role in biological fitness), whereas 'illnesses' are evaluative (i.e., *harmful* diseases that are judged to be undesirable, worthy of special treatment, and a valid excuse for normally criticizable conduct). Boorse (1997, 2014) subsequently reformulated this concept of 'illness' as practical 'disease-plus' concepts (e.g., 'treatable disease,' 'disabling disease') that require value-judgments. Wakefield (1992) argues that the concept of mental disorder involves a factual component (i.e., the failure of a mental mechanism to perform its naturally selected function) and an evaluative component (i.e., the condition is harmful by current cultural standards). Thus, Boorse and Wakefield maintain that judgments of function and dysfunction are value-free (and constitute the objective basis of mental disorders), while judgments of harm are value-laden. Against this view, Gagné-Julien argues that judgments of function are value-laden. Since Gagné-Julien maintains that values are present in decisions regarding *which mental capacities* are relevant for mental health, it is not clear why this evaluative aspect of mental disorders cannot be accounted for at the level of harms. For example, one could defend a hybrid account that maintains that genuine mental disorders are: (1) caused by biological dysfunction (e.g., the disruption of a CR function that currently contributes to biological fitness), and (2) harmful because *culturally-valued* mental capacities (e.g., attentional capacities, reading capacities) are disrupted. It is not clear why we should accept Gagné-Julien's assertion that judgments of function (rather than harm) are value-laden.

For the purposes of articulating a definition of mental disorder, one theoretical advantage of locating value-judgments at the level of harms—rather than the function of mental capacities—is it would yield a more restrictive account of mental disorder that distinguishes between clinically insignificant dysfunctional traits (e.g., *suboptimal* emotion regulation) and harmful dysfunctional traits (e.g., depression that significantly impairs an individual's ability to lead a normal life). By contrast, Gagné-Julien's strategy of inferring dysfunctional traits from a theory of *culturally-valued mental capacities* (e.g., happiness, intelligence, confidence) runs the risk of producing an overinclusive definition of mental disorder that renders clinically insignificant mental traits (e.g., suboptimal intelligence) dysfunctional. In the clearest indication of specific mental capacities that are relevant for mental health, Gagné-Julien suggests parenthetically that "regulation of emotion, perception, social processes, or even 'happiness' and well-being'" are plausible candidates. Boorse (1976) argues against the strategy of identifying

highly valued mental traits (e.g., happiness or well-being) with mental health (pp. 68-70). For Boorse, the appropriate contrast class for 'disease' should be *normal* functioning, not *ideal* functioning. If the proper contrast class for mental disorders (i.e., 'mental health') is *normal mental functioning*, then locating value-judgments at the level of functions (rather than harm) risks conflating normal functioning with ideal functioning. This reinforces the idea that—in addition to satisfying Alexandrova's social rules—Gagné-Julien's normative account of function needs to be constrained further by specifying the higher goals (e.g., leading a normal life) that healthy mental capacities (e.g., perception, mood regulation) are directed towards.

I have elsewhere expressed pessimism regarding the prospects of defining mental disorder in terms of dysfunction and argued that mental disorders should be conceptualized as *harmful biological kinds* (Tsou, 2021). This hybrid definition liberalizes the naturalistic (i.e., factual) component of mental disorder away from biological dysfunction to *biological kinds*: classes of abnormal behavior whose characteristic signs are constituted by a set of stable biological mechanisms. The normative (i.e., evaluative) component demands that the effects of a biological kind are *harmful* (i.e., compromise individuals' capacity to lead a normal or unimpeded life as judged by cultural standards). This account is motivated by various pragmatic considerations. A practical disadvantage of the accounts of biological dysfunction defended by Boorse and Wakefield is that they present naturalistic criteria that are difficult to apply (unambiguously) to candidate disorders. By contrast, the more deflationary requirement of biological kinds offers a naturalistic standard that is readily ascertainable in practice. While the requirement of biological kinds is sufficiently broad to include disorders (e.g., schizophrenia, bipolar disorder) caused by biological dysfunction as a subclass, it would also include *normal psychological reactions to stress* (e.g., acute depression or anxiety) that are underwritten by biological mechanisms that fall within the *normal range of biological functioning.* If the effects of some psychological reactions are *harmful* for individuals, ruling out these conditions as genuine mental disorders because they fail to satisfy a technical definition of 'dysfunction' would be pragmatically indefensible. Accordingly, the naturalistic standard of being a biological kind is sufficiently liberal to accommodate the broad range of conditions that mental health professionals treat. Conversely, it is sufficiently restrictive insofar as it demands that particular mental disorders share a *common biological basis*, which ensures that *classifications* of such kinds yield *projectable inferences* (i.e., reliable predictions about kind members on the basis of induction). Compared to the broad and ambiguous definition of mental disorder adopted in the DSM (APA, 2013, p. 20), which invokes the cryptic concept of 'dysfunctional psychological, biological, or developmental processes,' the criterion of biological kinds provides a more transparent and applicable naturalistic standard for deciding whether a condition should be included in the DSM or excluded from it. Given the large number of dubious diagnostic categories (e.g., 'histrionic personality disorder,' 'voyeuristic disorder,' 'intermittent explosive disorder,' 'dependent personality disorder') that one currently finds in the DSM, it is evident that the DSM requires greater clarity and transparency regarding what it aims to classify. Although I am skeptical that developing a normative theory of function is a promising strategy for ruling out such artificial disorders, the social account of objectivity defended by Gagné-Julien usefully highlights correctable problems in the DSM revision process.

References

Alexandrova, A. (2017). *A Philosophy for the Science of Well-being*. Oxford: Oxford University Press.

Alexandrova, A. (2018). Can the science of well-being be objective? *British Journal for the Philosophy of Science*, *69*(2), 421-445.

Amundson, R., & Lauder, G. V. (1994). Function without purpose: The uses of causal role function in evolutionary biology. *Biology & Philosophy*, *9*(4), 443-469.

APA (2013). *Diagnostic and Statistical Manual of Mental Disorders, 5th ed.: DSM-5*. Washington, DC: American Psychiatric Association.

Boorse, C. (1976). What a theory of mental health should be. *Journal for the Theory of Social Behaviour*, *6*(1), 49-68.

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science*, *44*(4), 542-573.

Boorse, C. (1997). A rebuttal on health. In J. M. Humber & R. F. Almeder (Eds.), *What is Disease?* (pp. 1-134). Totowa, NJ: Humana Press.

Boorse, C. (2002). A rebuttal on functions. In A. Ariew, R. Cummins, & M. Perlman (Eds.), *Functions: New Essays in the Philosophy of Psychology and Biology* (pp. 63-112). Oxford: Oxford University Press.

Boorse, C. (2014). A second rebuttal on health. *Journal of Medicine and Philosophy*, *39*(6), 683-724.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy*, *72*(20), 741-765.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Gagné-Julien, A.-M. (2021). Dysfunction and the definition of mental disorder in the DSM. *Philosophy, Psychiatry, & Psychology*.

Griffiths, P. E. (1993). Functional analysis and proper functions. *British Journal for the Philosophy of Science*, *44*(3), 409-422.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In T. S. Kuhn, *The Essential Tension: Selected Studies in Scientific Tradition and Change* (pp. 320-339). Chicago: University of Chicago Press.

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science*, *56*(2), 288-302.

Neander, K. (1991). Functions as selected effects: The conceptual analysts' defense. *Philosophy of Science, 58*(2), 168-184.

Tsou, J. Y. (2021). *Philosophy of Psychiatry*. Cambridge: Cambridge University Press.

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