Designing AI for mental health diagnosis: challenges from sub-Saharan African value-laden judgements on mental health disorders

Edmund Terem Ugar [1,2] Ntsumi Malele [1,3]

[1]Philosophy, University of Johannesburg, Auckland Park, South Africa
[2]Centre for Africa-China Studies, University of Johannesburg, Auckland Park, Gauteng, South Africa
[3]Centre for the Philosophy of Epidemiology, Medicine, and Public Health, University of Johannesburg, Auckland Park, Gauteng, South Africa

Correspondence to
Edmund Terem Ugar, Philosophy, University of Johannesburg, Auckland Park, South Africa; teremedmund@gmail.com

Received 6 November 2023
Accepted 10 February 2024
Published Online First 19 February 2024

To cite: Ugar ET, Malele N. *J Med Ethics* 2024;**50**:592–595.

ABSTRACT

Recently clinicians have become more reliant on technologies such as artificial intelligence (AI) and machine learning (ML) for effective and accurate diagnosis and prognosis of diseases, especially mental health disorders. These remarks, however, apply primarily to Europe, the USA, China and other technologically developed nations. Africa is yet to leverage the potential applications of AI and ML within the medical space. Sub-Saharan African countries are currently disadvantaged economically and infrastructure-wise. Yet precisely, these circumstances create significant opportunities for the deployment of medical AI, which has already been deployed in some places in the continent. However, while AI and ML have come with enormous promises in Africa, there are still challenges when it comes to successfully applying AI and ML designed elsewhere within the African context, especially in diagnosing mental health disorders. We argue, in this paper, that there ought not to be a homogeneous/generic design of AI and ML used in diagnosing mental health disorders. Our claim is grounded on the premise that mental health disorders cannot be diagnosed solely on 'factual evidence' but on both factual evidence and value-laden judgements of what constitutes mental health disorders in sub-Saharan Africa. For ML to play a successful role in diagnosing mental health disorders in sub-Saharan African medical spaces, with a precise focus on South Africa, we allude that it ought to understand what sub-Saharan Africans consider as mental health disorders, that is, the value-laden judgements of some conditions.

INTRODUCTION

Machine learning (ML) technologies have become prevalent in healthcare and/or medical decision-making, and clinicians are increasingly relying on these technologies for clinical diagnosis and prognosis of medical conditions.[1] The use of ML techniques has been applied to the diagnosis of neurodegenerating diseases such as Parkinson's, Alzheimer's, mild cognitive impairment,[2] cardiovascular diseases and skin diseases such as skin cancers.[3] In the domain of mental health, the use of ML to diagnose mental disorders has become increasingly ubiquitous. Some researchers have shown the increasing use of ML in the diagnosis of mental illness and the successes that these technologies have achieved in this domain.[4–9] Grünerbl *et al*[7] expose how smartphones can be used to monitor mental health disorders like bipolar. In addition, using ML features such as random forest, Reddy *et al*[8] show, in their research, how the system can predict the causes of stress in employees. Giannakakis *et al*[9] used ML techniques and heart rate variability to show the causes of stressors in selected patients. Recently, Iyortsuun *et al*[6] showed how deep learning, a subfield of ML, has been used to diagnose schizophrenia, and the technology obtained excellent accuracy. To diagnosis mental disorders, ML analyse patterns in patients' healthcare datasets collected from sources like brain imaging scans, social media posts, and patients' healthcare records.

While ML is achieving some success in diagnosing mental health diseases, we ask the following question: (a) given that what constitutes diseases, as we conceive in this paper, are explained from both a value perspective and a biostatistics/naturalist viewpoint, can there be a generic ML technology for diagnosing mental health disorders? In other words, can a universal/generic ML system be efficient in diagnosing mental health disorders, and if not, what is needed for ML designers to effectively design and implement effective ML systems in different climes while considering contextual variations of disorders?

In this paper, we contend that the 'successes' and accuracies of ML in diagnosing mental disorders will depend on both the value judgements of what a particular clime considers a mental disorder and on some naturalist explanations of disorders. As a result, a generic or universal design cannot be effective given the heterogeneity of value judgements in defining what mental health disorders are in different contexts. For instance, we claim that there are some possible challenges when it comes to successfully applying ML systems designed elsewhere within the sub-Saharan African context, especially in diagnosing mental health disorders. This is because mental health disorders cannot be diagnosed solely on 'factual evidence' but on both factual evidence and value-laden judgements of what constitutes mental disorders in sub-Saharan Africa. We point out some mental health cases that may be conceived as schizophrenia (caused by brain dysfunction) in South Africa by non-South Africans. Yet, South Africans see these conditions differently, not as an illness caused by the brain but as a spiritual possession. Given cases like the above, we adumbrate that for ML systems to be successful and accurate in diagnosing mental disorders in Africa, the technology and its designers must understand the value-laden judgement of what is considered a mental disorder within the African context to avoid misdiagnosis.

We begin this paper by first clarifying what we consider as diseases, precisely, mental health diseases. We engage the naturalist, normativist and hybrid accounts on mental diseases to enable us to situate the problem an ML technology may encounter when diagnosing mental disorders. Second, we engage some of the success in using ML to diagnose mental disorders. We argue that despite the success of ML, we draw some of the cases discussed in the first section to show the challenges ML systems might face in sub-Saharan

Africa. We conclude by making some practical recommendations that ML designers should bear in mind when designing ML systems for clinical diagnosis in healthcare.

INTERROGATING THE CONCEPTION OF MENTAL DISORDERS: THE NATURALISTS, NORMATIVISTS AND THE HYBRID ACCOUNTS

In the philosophy of medicine debate, mental health and illness can be divided into three main groups: naturalism, normativism and hybrid accounts of health. This debate questions whether certain mental illnesses should be classified as natural kinds, social constructs or a hybrid of both. Christopher Boorse[10] is the best-known proponent of a naturalist, purely biological account of disease. In his work, Boorse[11] argues that diseases are biological dysfunctions and can be determined independent of value judgements. Boorse presents a biostatistical theory to demonstrate that a condition qualifies as a disease solely based on value-free scientific facts. Boorse's biostatistical theory explains disease (or pathology in Boorse's term) as a statistically species-subnormal function informed by environmental factors where the species inhabit. To understand BTS, it is imperative to understand the following terms:

> (1) The reference class is a natural class of organisms of uniform functional design, specifically, an age group or the sex of a species. (2) A normal function of a part or process within members of the reference class is a statistically typical contribution to their individual survival and reproduction. (3) A disease is a type of internal state which is either an impairment of normal functional ability, that is, a reduction of one or more functional abilities below typical efficiency, or a limitation on functional ability caused by the environment. (4) Health is the absence of disease.[11]

According to this biostatistical theory, assessing an individual's health status involves comparing it to the health of a specific population. For a person to be deemed healthy, Boorse says that their bodily functions should fall within the statistically defined range of normality. However, if there is a deviation from this range, this indicates a state of ill health.[10] Kendell[12] also makes a similar argument. As a proponent of the naturalistic account, Kendell advances the notion that a disease is something that is value-free and based purely on scientific facts. Kendell[12] claims that there is sufficient evidence to support the notion that (psychotic mental illnesses such as) schizophrenia and manic-depressive illness possess an inherent biological disadvantage, which justifies categorising them as illnesses.

However, normativists challenge the naturalistic view by presenting cases that highlight the dynamic nature of disease classification.[13] Normativists contend that there are cases where a condition presents a clear biological dysfunction, but the condition cannot be categorised as a disease. For example, a man who undergoes a vasectomy could be considered diseased from the naturalist perspective. This is on the grounds that the individual's body function is not uniform with the rest of the natural class or the sex of the species (male). For instance, the individual would no longer contribute towards reproduction, and their reproductive organs would no longer function as others in the same reference class. However, from the normativist perspective, such conditions cannot be considered a disease. Why so? We briefly espouse this argument below.

The normativist explanation of diseases adequately addresses similar instances like the above by considering the alignment between shifts in disease classification and the underlying values of those who suffer from such conditions.[13] According to normativism, the terms 'health' and 'disease' reflect our values, as acknowledged by both lay people and medical professionals.[13] Normativists contend that 'healthy' describes the physiological or psychological states we desire, while 'diseased' is reserved for states we strive to avoid. Health and disease cannot be solely defined based on biological or evolutionary factors. Instead, these concepts are influenced by values and carry normative meaning. Similarly, Cooper[14] argues that the currently proposed naturalistic accounts are inadequate. Cooper claims that biological dysfunction does not necessarily result in disorder, meaning some biological dysfunctions are not disorders, like the case of vasectomy that we presented above. Thus, understanding health and disease requires recognising and examining the implicit values of biomedical definitions.

Normativists emphasise that social and political values shape our understanding of diseases; that is, in determining what qualifies as dysfunctional behaviour, normativists stress the significance of considering values. For instance, Cooper presents three requirements for a condition to be considered a disease. First, the condition must be bad for the individual who suffers from it; second, it has to make the individual unlucky; lastly, the condition has to be potentially medically treatable. For instance, if one who undergoes vasectomy does not think it is bad for them, or they do not think it makes them unlucky, such a condition should not be considered a disease, even though the vasectomy can be potentially reversed.

Although Cooper presents a convincing argument, we argue that just as scientific facts alone are insufficient in the naturalistic account of health and disease, relying solely on values is also inadequate for defining health and disease. The normativists theory is necessary for explaining mental health disorders; however, it is not sufficient on the grounds that there has to be a presence of a biological dysfunction to provide a plausible causal explanation of a condition as a disease, even from a normativist perspective. Furthermore, our friend who undergoes a vasectomy only knows if they feel unlucky or bad for them based on the general functioning of their reference class. In other words, a plausible explanation of diseases has to encompass the notion of biological dysfunction and the social values of one who bears the condition. As a result, we favour a hybrid account of health and disease that best explains mental health and grounds our conceptual framework.

Wakefield,[15] a proponent of the hybrid model of disorders, integrates both naturalist and normativist theories of health and disease. Wakefield emphasises a factual component within the concept of disorder to differentiate it from other disvalued conditions. However, facts alone are not enough. He adds that disorder requires harm, and that involves values. Therefore, for the concept of disorder, we need both values and facts. According to Wakefield,[15] a disorder is characterised as a 'harmful dysfunction,' where 'harmful' is a vague term based on societal norms, and 'dysfunction' is a purely scientific term. For Wakefield (1992: 383)[15], dysfunction signifies an unfulfilled function, indicating a failure of an organism's mechanism to perform its intended purpose. He suggests that a disorder is distinct from a failure to function in a socially desirable way because dysfunction is present only when an organ cannot naturally perform its function. However, according to Wakefield, a dysfunction is insufficient to justify attributing a disorder. Wakefield alludes that a disorder must have context-specific implications; that is, the sufferer must see the harm within their cultural contexts. This argument is in line with Cooper's conditions for a dysfunction to be considered a disorder. For instance, if the country of our friend who undergoes a vasectomy is overpopulated and needs to cut down on the birth rate to manage available resources, it follows that their decision to undergo a vasectomy has to be understood from the lens of their country.

To narrow the argument to mental health, Wakefield claims that a mental disorder is a harmful mental dysfunction. In his definition, Wakefield outlines the concept of mental disorder as follows:

> A condition is a mental disorder If and only if (a) the condition causes some harm or deprivation of benefit to the person as judged by the standards of the person's culture (the value criterion) and (b) the condition results from the inability of some mental mechanism to perform its natural function, wherein a natural function is an effect that is part of the evolutionary explanation of the existence and structure of the mental mechanism (the explanatory criterion).[15]

Wakefield argues that the concept of disorder imposes two requirements on theories of mental disorder. The first requirement, the value criterion, suggests that a successful theory must align with the commonly accepted notion of harm. In other words, it should explain how a condition or dysfunction causes harm to the individual according to that individual's societal standards.[15] The second requirement, the explanatory criterion, states that the physiological subsystem must fail to perform the function that explains that trait's persistence through the evolutionary process. Essentially, the theory should clarify and describe the underlying dysfunctions that contribute to the development of mental disorders.

Although Wakefield's account originates from a Western perspective, we believe that his hybrid approach to harmful dysfunction has the potential to incorporate theories of disorder even from an African standpoint. According to this approach, for a condition to be classified as a disorder, it must satisfy both the explanatory and value criteria. In various Southern African traditions, some conditions fulfil either the value or explanatory criterion but are not regarded as mental illnesses. For example, some individuals may exhibit symptoms of psychosis that naturalists would classify as a biological dysfunction. These symptoms might include 'disorientation, changes in consciousness, forgetfulness, amnesia, seizures, dissociation, delusions or alterations in behaviour and personality'.[16] However, in traditional African perspectives, these symptoms hold different meanings compared with dominant Western perspectives. Austine *et al* explain the symptoms in the following manner:

> Ukuthwasa: the calling to be a traditional healer whose role is to help in the caring and healing process of the whole person, addressing the condition and also restoring balance by addressing the alleged causes of the illness. Ukufa kwabantu: which broadly means 'the illnesses of the people' and is not recognised by the patient as an illness or disease per se, but as disturbance caused by breaking a taboo or displeasing an ancestor, or due to a spell that has been cast on the patient. These include Indiki, ufufunyane or izizwe (spirit possession), Umnyama (pollution or contamination), Ubuthakathi (bewitchment or sorcery), Ukuphonsa (curses or spells), Umkhondo/umeqo/idliso (some form of poisoning)[16]

Therefore, psychotic symptoms are frequently not considered mental disorders in an African context. Instead, they hold significant spiritual significance in many African communities. Ukuthwasa is not considered a mental illness in the Xhosa culture of South Africa, but some of its symptoms, including impairment, are similar to those of amafufunyana (evil spirits).[17] Ukuthwasa is training that one undergoes to accept the calling to serve their ancestors. If one resists the 'calling', that may lead to an illness (ukuphambana). However, such illnesses do not have the brain to play a causal explanatory role but the spirits through a spiritual possession.[i] Given these differences in understanding, in what follows, we contend that it is pertinent that an ML system for diagnosing mental health disorders is 'cognisant' of these differences if it is deployed in contexts like sub-Saharan Africa. This is because the technology might categorise some conditions as pathological while failing to 'understand' the value judgements of those conditions from sub-Saharan African perspectives. Why so? Because ML technologies are not value-neutral—they are designed to replicate the values, norms, and worldviews of their designers, and this is the same as ML techniques used for diagnosing mental health disorders.

ML SYSTEMS IN MENTAL HEALTH DIAGNOSIS: POSSIBLE CHALLENGES

As earlier stated, there has been an increased application of ML techniques in healthcare. Clinicians are now encouraging and emphasising the application of ML systems in medical decision-making.[1 3] ML is a subfield of artificial intelligence characterised by two broad components: the ubiquity of big data (electronic healthcare records (EHR)—in healthcare) and deep neural networks (DNN).[1 18] DNNs label objects in healthcare images more accurately than humans, given the available pool of EHR. Furthermore, DNN distinguishes data points based on how data flows through its layers and how iteratively warped the input space becomes.[1 18]

Given the capabilities of ML's DNN in interpreting data points and making predictions, ML has been efficient in making predictions in healthcare. The application of ML in healthcare is making healthcare decision-making, clinical diagnosis, and treatment more reliable and quicker. Topol[19] discussed that clinical decision-making done by clinicians is currently susceptible to flaws, such as cognitive biases and diagnostic errors. As a result, ML has the potential to mitigate these issues and enhance healthcare decision-making capabilities. In addition, besides being of importance to individuals to have a technology that is not prone to the diagnostic errors of human clinicians but that is accurate and efficient in its diagnostic and predictive abilities, ML has the potential to cut healthcare costs.

In the mental health domain, the use of ML to diagnose mental diseases has become increasingly prevalent. Some researchers have shown how ML diagnoses mental disorders.[4–6 20–22] For mental health to be diagnosed, a series of psychiatric interviews which cover suspected symptoms, the psychiatric history of the patients, and the patient's psychiatric examination are carried out. Furthermore, psychiatrists include psychological assessment tools that help identify the psychiatric symptoms of the patients.[6 23] Iyortsuun *et al*[6] show that the applications of ML, robotics and DNN in mental health diagnosis have demonstrated some form of success. The above approaches and techniques can identify the causes of certain mental illnesses, diagnose these illnesses and then predict possible treatment remedies and outcomes for the disorders.

Katarya and Maan[24] used ML methods to find the possible influence of mental health challenges on employees of non-technical companies by using data from the Open Source Mental Illness, a non-profit organisation that was conducted in 2019. Furthermore, during the COVID-19 pandemic, the above-mentioned researchers, in another study, identified predictors of psychological distress in individuals using ML technologies. In addition, Srinivasagopalan *et*

[i][1]We thank one of the anonymous reviewers for calling our attention to this aspect.

al[25] combined functional MRI and ML algorithms such as logistic regression, support vector machine and random forest to diagnose patients with schizophrenia. To modify their method, they applied feature selection to improve the model accuracy. In addition, the researchers then used a 'deep learning binary classifier including three hidden layers'. Their results showed higher accuracy than other traditional ML approaches mentioned above in diagnosing schizophrenia.[5]

While the use of ML has been somewhat efficient in diagnosing mental health diseases, we hold that given the exposition of mental health disorders, which we adumbrate in the first section, current designs of ML, given their non-value neutrality, may face some challenges if used for diagnosing mental health, especially in sub-Saharan Africa. This is because the technology may overgeneralise what it conceives as mental diseases, leading to misdiagnosis in the above context. This is because what is valued as diseases from dominant Western viewpoints, where most of these technologies are designed, may not be diseases in sub-Saharan Africa. Sub-Saharan Africans may not value some 'disorders' as diseases. However, the technology might not understand these differences and variations, given that the training datasets may lack some of these intricacies and worldviews.

For example, using the South African cases of ukuthwasa, ukufa kwabantu and ufufunyane, we argue that individuals who go through these conditions display the same symptoms as those who exhibit psychosis or schizophrenia. For example, psychosis or schizophrenia patients experience disorientation, delusion, personality alteration and changes in consciousness, the same as those who are experiencing the conditions mentioned above. It is most likely that ML systems designed to diagnose conditions like schizophrenia and psychosis may categorise individuals with the symptoms mentioned above as suffering from these mental health disorders caused by brain dysfunction when, from their cultural perspective, this condition might be an 'illness,' with a different causal explanation—spiritual rather than physical (brain). From an African cultural perspective, the symptoms may signify the presence of spiritual involvement. It is no fault of the technology itself to misdiagnose these individuals because it does not understand the above conditions. However, our intention in this paper is to draw the attention of designers to these normative differences in conceptualising mental health disorders and to point out that it is imperative that these differences are understood, given that ML designs are not value neutral.[26 27] To best design ML to diagnose mental health, we make the following recommendations.

RECOMMENDATIONS AND CONCLUSION

Having outlined the following problems that ML may face if employed to diagnose mental illnesses in sub-Saharan Africa, using the case of South Africa, we make the following recommendations for ML and DL designs. First, designers should avoid designing generic ML systems for mental health diagnoses. Before designers of these technologies carry out their designs, they must understand the context in which they are designing for and those they aim to benefit. Second, designers and those who deploy these technologies must understand the value judgement of the conditions in which they are designing or deploying their technology. ML systems should not be designed with a generalised perception of mental disorders. The above recommendation is to ensure that designers do not design the right technology for the wrong context or design biased and discriminatory ML for healthcare. Finally, given that Africa is a big market for healthcare technologies, designers should study the worldviews of Africans regarding diseases and craft their designs based on their understanding. Alternatively, Africans should carry out research on how to design ML to diagnose conditions like ukufa kwabantu and other healthcare concerns to mitigate possible issues like misdiagnosis.

Contributors Both authors contributed equally. ETU contributed to the section on machine learning, while NM contributed to the section on mental health. Both authors revised and edited the paper together.

Funding No Funding

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information. NA.

ORCID iD

Edmund Terem Ugar http://orcid.org/0000-0003-3034-5045

REFERENCES

1 Grote T, Berens P. On the ethics of Algorithmic decision-making in Healthcare. *J Med Ethics* 2020;46:205–11.
2 Myszczynska MA, Ojamies PN, Lacoste AMB, *et al*. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol* 2020;16:440–56.
3 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
4 Garriga R, Mas J, Abraha S, *et al*. Machine learning model to predict mental health crises from electronic health records. *Nat Med* 2022;28:8.
5 Melchionna M. Can machine learning, Wearable Tech help treat mental health?; 2023.
6 Iyortsuun NK, Kim S-H, Jhon M, *et al*. A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare* 2023;11:285.
7 Grünerbl A, Muaremi A, Osmani V, *et al*. Smartphone-based recognition of States and state changes in bipolar disorder patients. *IEEE J Biomed Health Inform* 2015;19:140–8.
8 Reddy US, Thota AV, Dharun A. Machine learning techniques for stress prediction in working employees. In: *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC*. IEEE, 2018: 1–4.
9 Giannakakis G, Marias K, Tsiknakis M. A stress recognition system using HRV parameters and machine learning techniques. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW*. Cambridge, United Kingdom: IEEE, 2019: 269–72.
10 Boorse C. A rebuttal on health. in: Humber, J., Almeder, R., Eds. What is disease? Totowa, NJ: Humana Press, 1997.
11 Boorse C. What a theory of mental health should be. *J Theory Soc Behav* 1976;6:61–84.
12 Kendell RE. The concept of disease and its implications for psychiatry. *Br J Psychiatry* 1975;127:305–15.
13 Ereshefsky M. Defining 'health' and 'disease. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and BIOMEDICAL Sciences 2009;40:221–7.
14 Cooper R. Psychiatry and philosophy of science. In: *Psychiatry and Philosophy of Science*. Routledge, 2014.
15 Wakefield JC. The concept of mental disorder: on the boundary between biological facts and social values. *Am Psychol* 1992;47:373–88.
16 Austine T, Bezuidenhout C, Botha K, *et al*. Abnormal Psychology- A South African perspective. 2nd edn. Cape Town: Oxford University Press, 2017.
17 Niehaus DJH, Oosthuizen P, Lochner C, *et al*. A culture-bound syndrome 'Amafufunyana' and a culture-specific event 'Ukuthwasa': differentiated by a family history of schizophrenia and other psychiatric disorders. *Psychopathology* 2004;37:59–63.
18 Buckner C. Deep learning: a philosophical introduction. *Philosophy Compass* 2019;14.
19 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
20 Molly G. Your next therapist could be a Chatbot App; 2022.
21 Gold A, Gross D. Deploying machine learning to improve mental health; 2022.
22 Marr B. AI in mental health: opportunities and challenges in developing intelligent Digital therapies; 2023.
23 Jencks SF. Recognition of mental distress and diagnosis of mental disorder in primary care. *JAMA* 1985;253:1903–7.
24 Katarya R, Maan S. Predicting mental health disorders using machine learning for employees in technical and non-technical companies. In: *2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE*. IEEE, 2020: 1–5.
25 Srinivasagopalan S, Barry J, Gurupur V, *et al*. A deep learning approach for diagnosing schizophrenic patients. JETAI 2019;31:803–16.
26 Ugar ET. The fourth industrial revolution, Techno-colonialism, and the sub-Saharan Africa response. *FITAJII* 2023;12:33–48.
27 Ugar ET. MA minor Dissertation. Johannesburg; 2022.