JOURNAL OF APPLIED ETHICS AND PHILOSOPHY

Center for Applied Ethics and Philosophy
Hokkaido University

vol.16

February 2025

ISSN 1883-0129

Journal of Applied Ethics and Philosophy

Printed in Japan

ISSN 1883 0129 (Print)
ISSN 1884 0590 (Online)

All queries should be directed to:
The Editor-in-Chief, Center for Applied Ethics and Philosophy
Faculty of Humanities and Human Sciences
Hokkaido University
N10 W7, Kita-ku, Sapporo 060-0810, Japan
jaep@let.hokudai.ac.jp

CONTENTS

Editorial Note

The *Journal of Applied Ethics and Philosophy* is an interdisciplinary journal that covers a wide range of areas in applied ethics and philosophy. It is the official journal of the Center for Applied Ethics and Philosophy (CAEP) at Hokkaido University. The aim of the *Journal of Applied Ethics and Philosophy* is to contribute to a better understanding of ethical and philosophical issues by promoting research in various areas of applied ethics and philosophy, and by providing researchers, scholars and students with a forum for dialogue and discussion on ethical and philosophical issues raised in contemporary society. The journal welcomes original and unpublished regular articles and discussion papers on issues in applied ethics and philosophy.

Nobuo Kurata
Editor-in-Chief

Many thanks to *Journal of Applied Ethics and Philosophy* reviewers

The *Journal of Applied Ethics and Philosophy* would like to thank the following individuals for generously reviewing manuscripts for us between February 2024 and January 2025. The support and expertise of these professionals promote and maintain the high quality of the journal's content. Thank you very much..

Andrew Komasinski
Jeffry Gayman

Relativism, Universalism and the Limits of the Pluriversal Model of Decolonization[1]

Olusola V. Olanipekun
Department of Philosophy, Obafemi Awolowo University

Abstract

The paper examines pluriversality as a model of decolonization and its implications for the decoloniality scholars. In the past few decades, the concepts such as colonization, post-colonization and decolonization have formed important vocabularies in the works of many decoloniality scholars who championed the emancipation of the knowledge production from Eurocentric episteme. For obvious reason, this is not unconnected to the colonial legacies and past experiences of the (ex-)colonized people. One of the major problems with the Eurocentric model of epistemic practice is that it mainly represents the view of the West going by its ideological roots hence, subjective and relativistic even though often packaged as objective and universal. Thus, for the decoloniality scholars, pursuing pluriversal agenda in the decolonization of knowledge in African academy and other non-western societies is often considered to be an alternative model. However, arising from the above decoloniality philosophy, the central question in this paper is that, how objective and universal is the suggested pluriversal model of decolonization? The paper contends that answer to this question is not as straight forward as the decoloniality scholars assumed due to the complexity of the model. The paper concludes by examining the implications of the complexity and the limits of the pluriversal epistemic model. .

Keywords: Colonization, Decolonization, Pluriversality, Epistemic Practice, Relativism, Universalism

Introduction:

The paper examines the viability of the pluriversal model of 'epistemic practice'[2] as a way of challenging the legacy of colonial rule. In recent times, there is a growing body of literature in the public domain about how the West colonized the non-west, and how the non-west attempt to decolonize themselves. In fact, the issue of decoloniality has become a perennial issue in the world of scholarship for the past few decades due to the persistence of colonialism in different guise with neocolonial labels.

Essentially, there are different ways by which scholars employed the concept 'decolonization'. One way by which Satya Mohanty conceived the idea of decolonization is that, it is the process of unlearning historically determined habits of privilege and privation of ruling and dependency on the colonizers.[3] This implies that certain imposed subjective way of life has been learnt, or passed on as a form of body of knowledge which requires unlearning due to lack of objectivity. The

1 This paper was first presented at the International Conference on "Decolonising the Humanities in the African Academy", organised by the Faculty of Arts, Obafemi Awolowo University, Ile-Ife, Nigeria, held between, 23-26th October 2022.

2 Epistemic practice is a concept associated with epistemology. As a concept, it is primarily related to knowledge as the frame of meaning within which people enact their lives. In a more recent time, epistemic practices as the socially organized and interactionally accomplished ways that members of a group propose, communicate, assess, and legitimize knowledge claim. See Cetina Knoor, *Epistemic Cultures: How the Science make Knowledge*, Cambridge: Harvard University Press, 2005. P.65. Inger Eriksson, "Enriching Learning Activity with Epistemic Practice" in *Nordic Journal of Studies in Educational Policy*, vol.1, no.1 (2016):2. Gregory J. Kelly, "Epistemic Practices and Science Education", in Peter Licona, ed., *Science: Philosophy, History and Education*, London: Springer, 2018, p.139

3 Satya P. Mogghanty, **"**Colonial Legacies, Multicultural Futures: Relativism, Objectivity, and the Challenge of Otherness", *PMLA*, Vol. 110, No. 1, Special Topic: Colonialism and the Postcolonial Condition (1995):110

above understanding leads to the question of relativism. Also, for Franz Fanon, decolonization is a historical process: that is to say it cannot be understood, it cannot become intelligible nor clear to itself except in the exact measure that we can discern the movements which give it historical form and content.[4] Fanon's position is important in the sense that it is imperative to understand the historical background or direction in which decolonization is being discussed. Is it in the area of politics, epistemic practice or language? For the purpose of this paper, decolonization is discussed primarily within the confine of epistemic practice. However, I may have cause to dabble into other areas such as politics and language.

Basically, research has shown that one of the main concerns of the decoloniality scholars is to devise a means of de-entrapping the non-west from every appearance of westernization. This leads to several calls to decolonize educational systems including methodologies for research among other things in the non-western societies. Arising from the fact that pluriversal model of decolonization of knowledge is often considered to be an alternative among the non-westerners, it should however be noted that little attention is often paid to the complexities of the model. The challenge is that, can the pluriversal model escape the problems that bedeviled the existing Eurocentric model of epistemic practice? Answer to this question shall determine the shape of this paper. It is important to note that this paper does not defend pluriversality neither does it defend western Eurocentric model. It mainly examines possible implications and problems with the highly celebrated pluriversal model of epistemic practice in the non-western academy.

The paper is divided into three main sections. Section one reflects on colonialism and the relativists challenge. This triggers our curiosity about the question of whether "man is really the measure of all things?" as argue by Protagoras, or "West is truly the measure of all things" as reflected by the domineering status of the western model of epistemic practice. Section two considers the challenge of Eurocentric-universalism. Section three, examines the complexities and the limits of the pluriversal epistemic model.

Colonialism and the Relativist Challenge: Is Man really the Measure of all Things or the West the Measure of all Things?

The Dialogue *Theaetetus* contains Protagoras *Homo Mensura* dictum. The dictum reads: "Of all things the measure is man: both of things that are (man is the measure) that they are, and of things that are not (man is the measure) that they are not".[5] This view is often interpreted to engender relativism. Arising from the above dictum, the issue of relativism has generated a serious discussion among scholars. When Protagoras argues that "man is the measure of all things" and the Western world decides the direction of all things including the direction of epistemic practice for the non-west, will it be illogical to restate Protagoras' dictum to read: "West is the measure of all things instead"? This may not be inappropriate given the Eurocentric approach to things. But in reality, should West be the measure of all things? The reason is because "man" in the above context as used by Protagoras implies all human beings both in the west and non-west.

However, is anything really wrong with the Eurocentric epistemic practice? The problem with the western epistemic practice is that it mainly represents the view of a particular set of people that is imposed on others. That is, the western Eurocentric model mainly represents the perspective of the western world alone. Thus, it is relativistic because the non-west/non-European were not part of its conception. The reason behind the relativist challenge is this. For instance, the western model of formal education or mode of acquiring knowledge was originally designed for the western people and not for Africa or other colonized societies. In other words, the point is that Africa as a continent was not in the agenda of the West when the western form of formal education was developed. The non-western societies also have their own existing form of education before the advent of the west. Where the problem lies is in the attempt to present the western model as the universal model that ought to be universally applied.

As corroborated by Tuck and Wayne, one noticeable trend is the ease with which the language of decolonization has been superficially adopted into education, supplanting prior ways of talking about social justice, critical methodologies, or approaches which decenter settler perspectives.[6] The influence of Western educational model in non-Western societies is frequently referred to as educational neocolonialism in the sense that Western paradigms tend to shape and influence educational systems and thinking elsewhere. Nguyen's position is that western epistemic practice tends to control and dictate the pace for non-west.[7]

Considering what is really wrong with western

4 Franz Fanon, The Wretched of the Earth*,* 1963, p. 36

5 Plato, *Protagoras*, Benjamin Jowett (trans.), Indianapolis: The Bobbs-Merrill Company, Inc., 1956, p. xii

6 Eve Tuck, and Wayne Yang K. "Decolonization is not a metaphor" in *Decolonization: Indigeneity, Education & Society*. Vol. 1, No. 1, (2012):2

7 Phuong-Mai Nguyen, Julian G. Elliott, Cees Terlouw, Albert Pilot, "Neocolonialism in education: Cooperative Learning in an Asian Context" *Comparative education*., vol. 45, no.1, (2009):109-130.

epistemic practice, Immanuel Wallerstein argues that, part of the legacy of colonialism was that educational development and standards of knowledge production in non-western academic environment are based on Western epistemological schema and theories that are deeply rooted in, and informed by, colonial thought.[8] Wallerstein's description of western epistemic practice in the above quotation reveals that it is narrow, imposing, and one-sided. In the same spirit, Tanika Sarkar added that the colonized nations have long been exposed to dominant western cultural-intellectual values which were re-presented to her as universally valid ones and as immensely superior to her own traditions.[9] The above scholars submitted that the idea of presenting western cultural- intellectual values as a universal epistemic practice is questionable. The reason is because, it is not really universal but a mere westernized idea that is presented as if it is universal.

The above view was what Tanika Sarkar was expressing when he argues that, western model of universalism has always been a counterfeit value which tries to ensure western cultural-intellectual domination of the non-west in the name of universal norms which, actually, derive from western traditions."[10] The problem with the western version of universalism is that it is not really universal in the real sense of the word, but often presented as one.

Similarly, Rolando Vazquez also pointed out that "The universal pretentions of modernity have functioned as mechanisms of exclusion... The affirmation of modernity's universalism, which is an expression of its total validity claims, is built on a double negation: the exclusion of the 'other'.[11] In line with the above view, Western paradigms tend to shape and influence educational systems and thinking elsewhere through the process of globalization. Given the perceived pressure to modernize and reform in order to attain high international standards, educational policy makers in non-Western countries tend to look to the West.[12] The above view by Nguyen is an expression of the major problem with the western version of universalism and western model of epistemic practice.

In decoloniality scholarship, research methodologies are never accepted as neutral but are unmasked as technologies of subjectivation if not surveillance tools that prevent the emergence of another-thinking, another-logic, and another-world view.[13] As Sabelo rightly pointed out, the above view implies that every research methodology is a product of a particular academic tradition. By implication, the West imposes their subjective research methodologies on the non-west. They tried to universalize it in an attempt to dominate other epistemic practices. This is an issue of a serious concern for the decoloniality scholars.

However, one of the western scholars who is sympathetic with decoloniality discourse is Sandra Harding. According to Harding, "Westerners must learn how to make ourselves fit, and to be perceived to be fit, to enter into the democratic, pluri-centric global dialogues from which global futures will emerge."[14] Post-colonial theorists allege that in the name of a universal moral order, their view has been plundered and imperialized by western cultural orientations during the colonial period.[15] In other words, the non-west accused the west of colonial imperialism. In essence, we can deduce that western universalism has always been a counterfeit value which tries to ensure western cultural-intellectual domination of the non-west in the name of universal norms which, actually, derive from western traditions. Now, arising from the fact that the coloniality scholars charged the Eurocentric universalism with the problem of relativism the next alternative is to think of pluriversality. What is pluriversal model? Answer to this question will be the main focus of the next section.

The Challenge of the Pluriversal Epistemic Practice in the Decolonial Discourse

Pluriversality refers to a particular value of a world in which many worlds fit.[16] In line with Robin Dunford's view, Escoba Arturo argues that Pluriversality

8 Çağrı Tuğrul Mart "British colonial education policy in Africa" in *Internal journal of English and literature*, Vol. 2, No. 9, (2011): 190-194.

9 Tanika Sarkar "How to Think Universalism from Colonial and Post -Colonial Locations: Some Indian Efforts" in Petter Korkman & Virpi Mäkinen (eds.), *Studies across Disciplines in the Humanities and Social Sciences, Helsinki: Helsinki Collegium for Advanced Studies*., 2008, p.240

10 Tanika Sarkar "How to Think Universalism from Colonial and Post -Colonial Locations: Some Indian Efforts" p.239

11 Rolando Vázquez, "Towards a Decolonial Critique of Modernity" in Raúl Fornet Betancourt (ed.), Capital, Poverty, Development, Denktraditionen im Dialog: Studien zur Befreiung und interkulturalität, Vol. 33, Wissenschaftsverlag Mainz: Aachen 2012, pp. 241-252

12 Nguyen, M. and Elliott, J. and Terlouw, C. and Pilot, A. "Neocolonialism in education: cooperative learning, Western pedagogy in an Asian context.", p.109

13 Sabelo J. Ndlovu-Gatsheni, "Decoloniality as the Future of Africa" in *History Compass,* vol.13, No.10, (2015): 489

14 Harding, Sandra. *Sciences from Below: Feminisms, Post-colonialities, and Modernities*. Durham, NC: Duke University Press, 2008. Also, see Shari Stone-Mediatore, "Global Ethics, Epistemic Colonialism, and Paths to More Democratic Knowledges" in *Radical Philosophy Review*, Volume 21, No. 2 (2018):1.

15 Tanika Sarkar, "How to Think Universalism from Colonial and Post-Colonial Locations: Some Indian Efforts", p.239

16 Robin Dunford, "Toward a Decolonial Global Ethics" in *Journal of Global Ethics*, vol. 13, no.3, (2017):391

implies a world where many worlds, worldviews and epistemologies fit. Pluriversality, denotes the existence of irreducibly plural ways of knowing and being that have survived the on-going violence of coloniality.[17] Similarly, Walter Mignolo is of the view that pluriversality implies the survival of myriad ways of knowing and being in the world that deny the authority of any knowledge system claiming universal validity or a transcendent grasp of 'objective' reality.[18] Also, pluriversality affirms the existence of 'multiple ontologies, multiple worlds to be known-not simply multiple perspectives on one world'[19]

Essentially, there are several ways by which the decoloniality scholars defined and defended the pluriversal model of epistemic practice. Garrett Fitzgerald started by analyzing the goal of the ontological aspect of pluriversality. According to Fitzgerald, the ontological aspect of pluriversality directly … consists of dismantling systems of power that threaten the survival of diverse ways of knowing and being.[20] In the context of our discussion, Garrett Fitzgerald argument is that there is an existing system of power defended by the western epistemic practice. However, contrary to such universal or western epistemic practice, pluriversal epistemic practice came to challenge and disprove such view with the claim that there are diverse ways of knowing.

For Phuong-Mai Nguyen, "non-Western cultures should seek to reconstruct imported pedagogic practices in accordance with their own world views and in line with their own norms and values".[21] This suggests a way of reconstructing epistemic practice in line with one's background. For Sardar Zed, the non-west has to create a whole new body of knowledge, rediscover its lost and suppressed intellectual heritage, and shape a host of new disciplines.[22] Essentially, pluriversality as a value suggests that practices, worldviews, values, or policies are legitimate only if they remain compatible with the existence of other worlds.[23] Thus, pluriversality sets a standard of legitimacy that would judge as morally wrong any worldview, value or practice that does not accept the existence of, or that works to shut down, other worlds, those holding such views ought not be excluded from dialogue.[24]

Still on pluriversal model, Shari Stone-Mediatore argues that pluriversality involves that "more responsible participation in global dialogue demands greater attention to the politics of our own knowledge-practices. We need to consider the kinds of attitudes and relationships that our own knowledge practices cultivate."[25] Essentially, because the universalizing discourse of modernity imperils the survival of other ways of knowing and being, embracing the ontological fact of pluriversality impels a corresponding rejection of epistemologies, discourses, and political projects that view the world as knowable and governable from within any single system of knowledge.[26]

Now, arising from the questions posed by Lisa Ausic, can a pluriversal epistemic practice truly foster a 'world where many worlds fit'? Is a pluriverse that continues to accommodate ontological dualism a pluriverse that the planet can bare?[27] These important questions are still searching for answers up till now. The reason is because, the decoloniality scholars do not pay attention to these questions but just busy romanticizing and over-celebrating pluriversality. Now, let us consider the challenge with the above view.

How Objective and Universal is the Pluriversal Model of Decolonization?

As hinted earlier, decoloniality scholars such as Mignolo and Escoba have argued that pluriversal model of decolonization devoid of relativism. Granted that this is possible in principle for the purpose of

17 Escobar, Arturo, *Designs for the Pluriverse: Radical, Interdependence, Autonomy, and the Making of the Worlds,* Durham: Duke University Press, 2018.

18 Mignolo, Walter, *The Darker Side of Western Modernity,* London: Duke University Press, 2011, pp.70-71

19 Conway Janet and Singh Jakket, *Radical Democracy in Global perspective: Notes from the Pluriverse*, Third World, Vol.32, No.4(2011):701.

20 Garrett Fitzgerald, "Pluriversal Peacebuilding: Peace Beyond Epistemic and Ontological Violence" *E-International Relations*, 2021, p.2

21 Phuong-Mai Nguyen, Julian G. Elliott, Cees Terlouw, Albert Pilot, "Neocolonialism in education: Cooperative Learning in an Asian Context" *Comparative education.*, vol. 45, no.1, (2009):109

22 Sardar, Z. Development and the locations of Euro-centricism. In R. Munck & D. O'Hearn, (Eds). *Critical development theory,* London: Zed, 1999, p.57.

23 Robin Dunford, "Toward a Decolonial Global Ethics" in *Journal of Global Ethics*, vol. 13, no.3 (2017):391

24 Robin Dunford, "Toward a Decolonial Global Ethics", p. 91

25 Shari Stone-Mediatore, "Global Ethics, Epistemic Colonialism, and Paths to More Democratic Knowledges" in *Radical Philosophy Review*, Vol. 21, No. 2 (2018):19-20

26 FiztGerald, Garrett, "Pluriversal Peacebuilding: Peace Beyond Epistemic and Ontological Violence." *E-International Relations*, 2021, p.9. Visit: https://www.e-ir.info/2021/11/27/pluriversal-peacebuilding-peace-beyond-epistemic-and-ontological-violence/violence/#:~:text=Because%20the%20universalizing,system%20of%20knowledge

27 Lisa Ausic (2022): Pluriversal Politics: The Real and the Possible, *Politics, Religion & Ideology*, Vol. 23, No.3(2022):1-3

argument, but in practice, can the pluriversal model of decolonization escape the relativist challenge? The answer to this question is not as straight forward as it is assumed. Several decoloniality scholars who suggested pluriversal methodology answered the question in the affirmative. However, a careful examination suggests that their response may not be appropriate. For instance, the essence of pluriversality is to produce a valid African or non-western model of Humanities that is universal and objective. But in what sense will it be universal? Will it be substantively universal or a form of relative universalism? If it will be substantively universal, how appropriate will that be, especially when the Western world does not believe that it is applicable to them? Also, if it will be relatively universal, it will run contrary to the initial position maintained by the decoloniality scholars such as Escoba that pluriversality devoid of relativism. From a critical point of view, someone may say that the African model of humanities, for instance, will mainly lead to perspectivism and not relativism, but from a careful observation, perspectivism is also an offshoot of relativism.

For Satya Mohanty, "philosophers agree that the "subjective" component of human knowledge is unavoidable, not because we are weak and erring creatures but because our epistemological criteria and our methodology are profoundly mediated by organized pre-suppositions and beliefs."[28] Mohanty's view attests to the fact that , the lure of relativism is especially strong when the justified and reasonable caution about ethnocentric idealizations of rationality and a narrow view of objectivity is inflated to a vague and undifferentiated skepticism toward knowledge.[29] The above view is well supported by Nicholas Sturgeon. According to Sturgeon, "Listening carefully to opposing views, regarding them as a challenge to one's own and attempting to appropriate their insights is, among other things, a mark of mutual respect that can provide a bond even across considerable disagreement."[30] According to Nicholas Sturgeon, "There is no requirement, as there is in the proposed transition to a relativistic stance, that one adopt either standard on the ground that it is one's own."[31]

For Sturgeon, the fallback stance recommended by relativism is not an understanding that tries somehow to incorporate what is right in all the competing perspectives. What it recommends instead is simply a retreat to reliance on one's own standards or on one's own reactions.[32] Given Sturgeon's comment above, I do not see how African model or Asian models of epistemic practice will escape the above relativists' challenge.

Similarly, for Dussel, a decolonial perspective does 'not presuppose the illusion of a non-existent symmetry between cultures.'[33] One thing about the pluriversal model is that it allows Africans to have their own model of epistemic practice. The question is, how universal and objective will African model of epistemic practice be? This is an important question that require urgent attention for the decoloniality scholars who championed pluriversality.

To support the above view, Mignolo argues that pluriversality is an option to be embraced by all those who will actively engage, politically and epistemically, to advance projects of epistemic and subjective decolonization and in building communal futures.[34] The implication of the above view is this. In an attempt to build an epistemic practice with communal features, an African model or Asian model of epistemic practices will not escape the same problem of relativism faced by the western model. Now, even if the African or the Asian models are universalized, this paper argues that the best they can be is relative-universal model, which is not different from the most criticized western model. An attempt to pursue this further leads us to the question of pluriversality and the problem of incommensurability that will be discussed in the next section.

Pluriversality and the Problem of Incommensurability

From all indication, it is obvious that pluriversality was conceived out of the believe in incommensurability thesis which is an extension of relativism. In the context of this discussion, incommensurability implies that there is no standard to judge whether one epistemic practice is better than another epistemic practice. That is, there is no standard to judge whether western model is better than non-western model. But if that is the case,

28 Satya P. Mohanty "Epilogue. Colonial Legacies, Multicultural Futures: Relativism, Objectivity, and the Challenge of Otherness" pp. 108-118

29 Satya P. Mohanty "Epilogue. Colonial Legacies, Multicultural Futures: Relativism, Objectivity, and the Challenge of Otherness", p. 112

30 Sturgeon, Nicholas L. "Moral Disagreement and Moral Relativism." in *Cultural Pluralism and Moral Knowledge*. Ed. Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul. New York: CUP, 1994. 113.

31 Nicholas L. Sturgeon, "Moral Disagreement and Moral Relativism" p.115

32 Nicholas L. Sturgeon, "Moral Disagreement and Moral Relativism" p.113

33 Dussel, Enrique. (2012) "Transmodernity and Interculturality: An Interpretation from the Perspective of the Philosophy of Liberation", *Transmodernity: Journal of Peripheral Cultural Production of the Luso-Hispanic World* 1 (3): 28-59, see Robin Dunford, "Toward a Decolonial Global Ethics" in *Journal of Global Ethics*, vol. 13, no.3 (2017):393

34 Mignolo, Walter, *The Darker Side of Western Modernity,* London: Duke University Press, 2011, p.27

will such position be a correct philosophical position to hold? According to Sardar, the non-Western cultures have to reconstruct themselves, almost brick by brick, in accordance with their own world views, norms and values by creating a whole new body of knowledge, rediscover its lost and suppressed intellectual heritage, and shape a host of new disciplines.[35]

Similarly, Ugo Zilioli argues that "All ancient and modern critics of relativism have suggested that the lack of an objective paradigm of measurement and commensuration opens the field to dangerous irrationality or to the use of force in place of reasoned persuasion."[36] Meanwhile for Grosfoguel, pluriversality is not a relativism of anything goes.[37] The implication of the above quotation is that Grosfoguel admits that pluriversality is not really free from the problem of relativism, it is only free from 'anything goes' version of relativism.

The Implications of the Pluriversal Model of Epistemic Practice in Non-Western Academy

One of the major arguments in support of pluriversal epistemic practice is that it fosters a world where many worlds fits. However, as pointed by Lisa Ausic, can a pluriversal epistemic practice truly foster a 'world where many worlds fit'? Is a pluriverse that continues to accommodate ontological dualisms a pluriverse that the planet can bare?[38] One important implication of pluriversal episteme practice is that, if there is a world that houses several other worlds, the question is what kind of world will that be? Such a world will be so complex than what we are used to. Imagine a world with western epistemic practice, African epistemic practice, Asian epistemic practice and so on. The question is that how objective and universal will these different epistemic practices be? How universal is Asian value? By decoloniality, it is meant here the dismantling of relations of power and conceptions of knowledge that foment the reproduction of racial, gender, and geo-political hierarchies that came into being or found new and more powerful forms of expression in the modern/colonial world.[39] The point of emphasis is that the unnecessarily multiplication of different worlds with different cultural outlooks and different epistemic practices within the main world will make the main world to be more complex than what it is.

Similarly, it is not clear how Afrocentric epistemic practice or Asian centered epistemic practice will be different from or escape the problems that bedeviled the Eurocentric epistemic practice. This point was corroborated by Barry Halen. For Halen, arguing that any non-Western system of cognition deserves an equal hearing, prima facie credibility in its own right as an alternative pathway to the "true" or "truth… There will be complaints that this kind of attitude opens the door to relativism and the loss of objective knowledge altogether.[40] Following the above view, Chaves Martha acknowledged that pluriversal politics or epistemic practice is not without challenges of its own. In the final analysis, it is important to understand that reflecting on pluriversality, caution must be taken against romanticizing the pluriverse as a place free from power or struggle.[41] The above view emphasized by Chaves, MacIntyre and Gerald is what this paper attempted to examine.

Conclusion

The paper examined the pluriversal model of decolonization and issues that surround it. Given that one of the major problems with the Eurocentric model of knowledge is that it mainly represents the view of the west going by its ideological roots, hence, subjective and relativistic even though often packaged as objective and universal. Thus, pursuing the Pluriversal agenda in the decolonization of knowledge in African Academy and other non-western societies is often suggested as an alternative model. However, arising from the above decolonial philosophy, the central question that runs through the paper was that, how objective is the suggested pluriversal epistemic practice? This paper discovered that the answer to this question is not as straight forward as the

35 Sardar, Z. "Development and the locations of Euro-centricism" in R. Munck & D. O'Hearn, (Eds). *Critical development theory,* London: Zed,1999, p.57

36 Ugo Zilioli, *Protagoras and the Problem of Relativism*, Hampshire: Ashgate,2007, p.77

37 Grosfoguel, Ramon. "Decolonizing Western Universalisms" in *Trans-modernity: Journal of Peripheral Cultural Production of the Luso-Hispanic World Vol.*1, No.3 (2012):101

38 Lisa Ausic (2022), "Pluriversal Politics: The Real and the Possible" in *Politics, Religion & Ideology*, p.1 3

39 Maldonado-Torres, N., 'Thinking Through the Decolonial Turn: Post-continental Interventions in Theory, Philosophy, and Critique—An Introduction', Trans-modernity: *Journal of Peripheral Cultural Production of Luso Hispanic World*, 1(2) Fall (2011):117

40 Barry Halen, *The Good, the Bad and the Beautiful*, Bloomington: Indiana University Press, 2000, p.35

41 Chaves, Martha, Thomas Macintyre, Gerard Verschoor, and Arjen E. J. Wals. 2016. "Towards Transgressive Learning through Ontological Politics: Answering the 'Call of the Mountain' in a Colombian Network of Sustainability. *Sustainability* 9 (1): 21. See Garrett Fitzgerald, "Pluriversal Peacebuilding: Peace Beyond Epistemic and Ontological Violence" *E-International Relations*, 2021.

decoloniality scholars assumed due to certain constraints that was discussed in this paper. The paper was concluded by considering certain implications of the model.

References

Audi, Robert., *Cambridge Dictionary of Philosophy*, Cambridge: CUP, 1999.

Ausic, Lisa., "Pluriversal Politics: The Real and the Possible" in *Politics, Religion & Ideology*, Vol. 23, No.3(2022):1-3

Çağrı Tuğrul Mart "British colonial education policy in Africa" in *Internal journal of English and literature*, Vol. 2, no. 9(2011): 190-194.

Chaves, Martha, Thomas Macintyre, Gerard Verschoor, and Arjen E. J. Wals. 2016. "Towards Transgressive Learning through Ontological Politics: Answering the 'Call of the Mountain' in a Colombian Network of Sustainability. *Sustainability.* Vol. 9, No.1, (2016):21. Conway, Janet and Singh Jakket, Radical Democracy in Global perspective: Notes from the Pluriverse, Third World, Vol.32, no.4(2011):689-706.

Dunford, Robin., "Toward a Decolonial Global Ethics" in *Journal of Global Ethics*, vol. 13, no.3 (2017):391

Dussel, Enrique. (2012) "Trans-modernity and Interculturality: An Interpretation from the Perspective of the Philosophy of Liberation", *Trans-modernity: Journal of Peripheral Cultural Production of the Luso-Hispanic World* 1 (3): 28-59

Escobar, Arturo, *Designs for the Pluriverse: Radical, Interdependence, Autonomy, and the Making of the Worlds,* Durham: Duke University Press, 2018.

Fitzgerald Garrett, "Pluriversal Peacebuilding: Peace Beyond Epistemic and Ontological Violence" *E-International Relations*, 2021, p.9. Also, visit; https://www.e-ir.info/2021/11/27/pluriversal-peacebuilding-peace-beyond-epistemic-and-ontological-violence/violence/#:~:text=Because%20the%20universalizing,system%20of%20knowledge

Franz Fanon, *The Wretched of the Earth,* Harmondsworth: Penguin,1963.

Grosfoguel, Ramon. "Decolonizing Western Universalisms" *Transmodernity: Journal of Peripheral Cultural Production of the Luso-Hispanic World Vol.*1, No.3 (2012):101,

Halen, Barry., *The Good, the Bad and the Beautiful*, Bloomington: Indiana University Press, 2000.

Harding, Sandra. *Sciences from Below: Feminisms, Post-colonialities, and Modernities*. Durham, NC: Duke University Press, 2008.

Inger Eriksson, "Enriching Learning Activity with Epistemic Practice" in *Nordic Journal of Studies in Educational Policy*, vol.1, no.1 (2016):2

Kallaway, Peter., "Welfare and Education in British Colonial Africa, 1918–1945" in Damiano W. O. Thompson, "Educational Objectives in The Modern College or University" *'Journal of Education*, (1924):13

Kelly, Gregory J., "Epistemic Practices and Science Education", in Peter Licona, ed., *Science: Philosophy, History and Education,* London: Springer, 2018.

Knoor, Cetina., *Epistemic Cultures: How the Science make Knowledge,* Cambridge: Harvard University Press, 2005.

Maldonado-Torres, N., 'Thinking Through the Decolonial Turn: Post-continental Interventions in Theory, Philosophy, and Critique—An Introduction', Transmodernity: Journal of Peripheral Cultural Production of Luso Hispanic World, vol.1 No.2, (2011):117

Matasci Miguel Bandeira Jerónimo, Hugo Gonçalves Dores, eds. Education and Development in Colonial and Postcolonial Africa, Cham: Palgrave Macmillan, 2020.

Mignolo, Walter, *The Darker Side of Western Modernity,* London: Duke University Press, 2011.

Nguyen, M. and Elliott, J. and Terlouw, C. and Pilot, A. "Neocolonialism in education: cooperative learning, Western pedagogy in an Asian context.", *Comparative Education.*, vol. 45, No.1, (2009):109

Phuong-Mai Nguyen, Julian G. Elliott, Cees Terlouw, Albert Pilot, "Neocolonialism in education: Cooperative Learning in an Asian Context" *Comparative education.*, vol. 45, no.1, (2009):109-130.

Plato, *Protagoras*, Benjamin Jowett (trans.), Indianapolis: The Bobbs-Merrill Company, Inc.,1956

Rolando Vázquez, "Towards a Decolonial Critique of Modernity" in *Raúl Fornet Betancourt (ed.), Capital, Poverty, Development, Denktraditionen im Dialog: Studien zur Befreiung und interkulturalität, Vol. 33, Wissenschaftsverlag Mainz: Aachen 2012, pp. 241-252*

Sabelo J. Ndlovu-Gatsheni, "Decoloniality as the Future of Africa" in *History Compass* vol.13, No.10, (2015): 489

Sardar, Z. "Development and the locations of Eurocentricism" in R. Munck & D. O'Hearn, (Eds). *Critical development theory,* London: Zed,1999, p.57

Satya P. Mohanty, "Colonial Legacies, Multicultural Futures: Relativism, Objectivity, and the Challenge of Otherness", *PMLA*, Vol. 110, No. 1, Special Topic: Colonialism and the Postcolonial Condition (1995):110

Shari Stone-Mediatore, "Global Ethics, Epistemic Colonialism, and Paths to More Democratic Knowledges" in *Radical Philosophy Review*, Vol. 21, No. 2 (2018):19-20

Sturgeon, Nicholas L. "Moral Disagreement and Moral Relativism." in Cultural Pluralism and Moral Knowledge. Ed. Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul. New York: CUP, 1994.

Tanika Sarkar "How to Think Universalism from Colonial and Post -Colonial Locations: Some Indian Efforts" in Petter Korkman & Virpi Mäkinen (eds.), *Studies across Disciplines in the Humanities and Social Sciences*, 4. Helsinki: Helsinki Collegium for Advanced Studies. (2008):240

Tuck, Eve and Wayne Yang K. "Decolonization is not a metaphor" in *Decolonization: Indigeneity, Education & Society. Vol.* 1, No. 1, (2012):2

Zilioli, Ugo., *Protagoras and the Problem of Relativism*, Hampshire: Ashgate,2007.

Discussion Paper:

The Limits of Explainability in Health AI - Why Current Concepts of AI Explainability Cannot Accommodate Patient Interests

Thomas Ploug[1,2] & Søren Holm[3,4]

[1] Centre for AI Ethics, Law, and Policy, Department of Communication and Psychology, Aalborg University

[2] School of Health and Welfare, Halmstad University

[3] Centre for Social Ethics and Policy, Department of Law, School of Social Sciences, University of Manchester

[4] Centre for Medical Ethics, HELSAM, Faculty of Medicine, University of Oslo

Abstract

In this paper we explicate the general concept of 'an explanation' and show that because there are many kinds of explanation there must be many kinds of 'explainability'. Subsequently we analyse the types of explanations we can give of Artificial Intelligence (AI) systems and their output using current explainability methods, and then discuss what types of explanation patients are likely to seek as part of the diagnostic process or as part of choice of therapy. We argue that the types of explanation that is provided by current AI explainability methods do not adequately answer many reasonable requests for explanation that patients can make when their diagnosis or treatment choice has involved the use of AI advice.

Keywords: Artificial Intelligence, Contestability, Explanation, Explainability

INTRODUCTION

Artificial intelligence (AI) systems for diagnosis and treatment choice are currently being introduced into routine use in a wide range of health care settings. (Rajpurkar et al., 2022) The visions for the future development and use of AI in health care are grand and it is claimed that AI will revolutionise health care and the role of health care professionals in the near future. (Topol, 2019) The use of AI systems is also a crucial component in the visions for precision and personalised medicine.(Bonkhoff & Grefkes, 2022; Nirvik & Kertai, 2022; Subramanian et al., 2020) Most of the AI systems being implemented and under development are based on a number of machine-learning architectures such as neural networks, support vector machines, random forest classifiers or similar.(Ferdous et al., 2020) These architectures are all algorithmic. When the AI model has been trained there is a determinate connection between the input and the output, and if the learning function has been turned off and the model 'locked' the same input will always produce the same output. Despite being algorithmic these AI systems are nevertheless 'black boxes' in the sense that it is almost always impossible

for human beings, even those who are domain experts to understand in detail how a certain input leads to a certain output. The systems are too complex and they may pick up on features in the input that have no readily available human interpretation, e.g. a certain pixel texture in a subsection of an X-ray. This is widely recognised as a problem. To simply quote the Little Britain catch phrase "Computer says No!" to a patient would obviously not be a good reason to deny that patient treatment if that is all that can be said. The most common response to this problem in the literature is to claim that AI systems in health care should be explainable, and there are very active research programs aimed at developing 'explainable AI' on the basis of the current AI architectures (see more below). A requirement for explainability is mentioned in many guidelines and policy documents, e.g. in the recent WHO guidance on AI in health.(World Health Organisation (WHO), 2021) A demand for explainability has also begun to appear in legislation and regulation especially in Europe. The European Union's General Data Protection Regulation Article 13(3)f requires the data controller to disclose "the existence of automated decision-making, including profiling referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, …".(General Data Protection Regulation, 2016) The requirement of provision of information about 'the logic involved' may charitably be interpreted to involve not only a general account, e.g. 'we process your information using a machine-learning support vector machine', but an individualised explanation of the processing of the data of the data subject. The EU Artificial Intelligence Act goes further and requires in Article 13 that "High-risk AI systems [which would include most or all AI systems in health care] shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately" whereas Recital 59 concerned with AI systems in law enforcement calls for such systems to be "explainable", and requires them to be defined as high-risk if they are not.(Artificial Intelligence Act (Regulation (EU) 2024/1689))

In the US the White House Office of Science and Technology Policy has issued a non-legally binding white paper entitled 'Blueprint for an AI Bill of Rights' that *inter alia* states that:

> "Automated systems should provide explanations that are technically valid, meaningful and useful to you and to any operators or others who need to understand the system, and calibrated to the level of risk based on the context."(White House Office of Science and Technology, 2022)

But is explainability the most important desideratum in a health AI system, alongside the (hopefully) *sine qua non* of precision and accuracy at a level equal or superior to that of a trained health care professional?

In this paper we will first explicate the general concept of 'an explanation' and show that because there are many kinds of explanation there must be many kinds of 'explainability' and a significant risk of equivocation when using these terms. We will analyse some of the kinds of explanation we are able to give of AI systems and their output using current explainability methods; and then discuss what types of explanation patients are likely to seek as part of the diagnostic process or as part of choice of therapy. Comparing the kind of explainability we can obtain in relation to AI systems with the explanations patients want and need will show that AI explainability falls short of meeting the reasonable requirements of patients. Based on previous work we will argue that some requests for explanation are better understood as examples of contestation of AI pronouncements on behalf of patients, and that answering such contestations may involve, but is not limited to explainability.

WHAT IS AN EXPLANATION?

When a philosopher hears the term 'explanation' the first example that springs to mind is probably Hempelian nomological-deductive or 'covering law' explanation in the philosophy of science where an observation is fully explainable if it can be deduced from a covering natural law. But explanation is a protean concept and there are many other types of explanation. From philosophy of science we also get the concept of a probabilistic explanation of an observation of a phenomenon, and from the philosophy of action a number of different explanatory schemata for explaining actions (e.g. in terms of goals, intentions, motives, compulsions etc.). In addition we have semantic explanations of the meanings of words and phrases, practical explanations of the functions or working of things, social explanations of the meaning or conventionality of social practices, justificatory explanations for apparent transgressions, existential explanations for the wonders and woes of being and many more. In each case there is an explanandum, something to be explained, and an explanans the thing (very broadly conceived) that explains the explanandum. Although all of these explanation schemata are therefore structurally similar, the character of the explananda and the appropriate explanantia, and therefore what counts as a good explanation, vary widely between cases. It is important to note that something can be a good explanation even if there are deeper and more basic explanations available. The explanation 'probably because you have 60 pack-years of smoking' is a perfectly good and adequate

explanation in response to the question 'doctor, why have I got lung cancer?' even though there is a more basic causal explanation in terms of carcinogens and mutational events in lung cells.

In a paper published in 1962 Passmore suggests that despite the wide variety of types of explanation there is nevertheless something that unites all proper explanations, and that is a lack of some specific knowledge and a sense of puzzlement or unfamiliarity in relation to that lack. To quote Passmore:

> "Everything depends, then, on what I know and what I want to know. If I am asked by an adult human being why Jones died, it will be no explanation to reply: 'All men are mortal', for so much, it can be presumed, he knows already; that is not the unfamiliar feature of the situation that is bothering him."(Passmore, 1962)

Passmore's analysis may be rephrased in terms of epistemic interests. Essentially his claim is that underlying all different types of explanations is a specific epistemic interest, i.e. an interest in acquiring some specific knowledge, and this interest determines what counts as a relevant explanation. The adult's interest in knowing why Jones died is an interest that renders general information on the mortality of all men irrelevant and picks out some set of specific information on Jones' death as relevant.

Because of the very large, possibly infinite number of different explanation types there is a risk of equivocation when talking about explanations, especially in relation to whether something is an explanation or whether an explanation has been provided. If a person asks for an explanation of a certain type and is given an explanation of another type it is at the same time true that an explanation has been given and that no explanation has been given to the person who asked for one since they are none the wiser in relation to their original query. To put it in terms of epistemic interests, the relevant person's epistemic interest has not been satisfied. This may happen even if the person offering the explanation does so in good faith intending to properly answer the request for an explanation.

In medicine this probably most often happens when a medical professional provides a technical explanation in relation to a question which was not technical or only partly technical. A patient having a contrast-enhanced ultrasound scan could potentially ask for an explanation along the lines of 'Why was contrast being used for my scan?'. In response a health care professional could offer an explanation something like 'the contrast more strongly reflects sound waves, and this can be seen on the scans as lighter areas. Since the contrast is in your blood the scan therefore reveals the bloodflow through your body and organs. Cancer and other kinds of diseased tissue usually have higher blood flow and therefore look lighter on the ultrasound image.' While such an explanation certainly answers the question posed by the patient, it is unlikely to satisfy the patient's potential, underlying interest in knowing 1) why it was used in his or her particular case, e.g. if there is a suspicion of a specific disease requiring contrast-enhanced ultrasound scans, 2) what are the pros of using contrast, e.g. that it increases the accuracy of the diagnostics, and 3) what are the cons of using contrast, e.g. what are the potential side-effects. The trained and experienced health care professional may obviously offer explanations that to a large extent accommodate the patients' interest. The example shows, however, that in medicine as elsewhere explanations can be provided that answers a specific request for an explanation but does not satisfy an underlying epistemic interest.

It follows from this analysis of the concept of explanation illustrating its protean nature that there will also be a large number of types of 'explainability' each referring to a specific type of explanation that is provided in response to a specific type of explainability question.

WHAT TYPE OF EXPLAINABILITY IS AI EXPLAINABILITY?

Because the current machine-learning AI structures are fully algorithmic and deterministic once learning mode is switched off and they are locked one type of explanation that could be provided would, if we take a neural network architecture as an example be a full mapping of input to neurons in the first layer with the weightings these neurons assigned to the input and transmitted to neurons in the second layer and up through the layers until we reached the output layer and the weightings of the neurons in that layer that generated the readable output, e.g. that there is a 77% chance that the patient has condition X, 10% chance that they have Y, and 13% chance that they have a condition that the network cannot classify. This would be a complete explanation, but it would not be helpful except in exceptionally simple cases, e.g. cases where the network had picked up on a patognomic feature of a condition (in which case there would be 100% certainty in the output), or where a very specific genetic variant made most of the difference. It would not be helpful because the weightings of the neurons would be uninterpretable, partly because of the amount of data, partly because many of them would have no readily available human interpretation.

The work in the research area of explainable AI is therefore aimed at generating simplified mappings of the 'reasoning processes'[1] employed by the AI

1 Reasoning processes in scare quotes because current(?) AI architectures do not have any mental representations of any reasoning processes.

system in calculating the output from the input. Most of the currently developed and widely implemented AI interpreters provide explanations of the output-input mapping at the system level, and not in relation to an individual case. Because the explanations are simplifications they may obscure important features of the actual workings of the system.(Ghassemi et al., 2021)

The probably most used method is the Local Interpretable Model – Agnostic Explanations – LIME method which can be applied to any type of machine-learning architecture. It works by performing multi-feature perturbations around a specific output and uses the result to fit a linear model incorporating feature importance and feature interactions. (PricewaterhouseCoopers (PwC), 2018; Ribeiro et al., 2016) A LIME model of a diagnostic system would for instance be able to provide an explanation of what features of the input, and what feature interactions are important when the system output is a diagnosis of a particular condition/disease.

More specific methods are available for particular architectures that can provide more accurate information about how the AI model uses input features to perform the classification task. In relation to Deep Neural Network (DNN) it is for instance possible to apply Relevance Propagation methods. A relevance propagation algorithm starts at the output and works backwards through the layers of the neural network, and assigns relevance scores to the inputs received from the preceding layer relative to a 'neutral' activation state of the network.(Shrikumar et al., 2017) These relevance scores provides relatively accurate explanations of the workings of the DNN, but may not always be readily explainable in human terms.

In relation to AI analysis of medical images a range of model specific methods have been used to generate 'heat maps' or saliency maps that visually indicate which areas of the image have been most important for the AI algorithm in classifying the image.(Rajpurkar et al., 2017; Selvaraju et al., 2017) This is supposed to aid the human interpreter, but may in certain circumstances be misleading because the heat map does not indicate which feature in the highlighted area that was important. (Ghassemi et al., 2021)

It follows from the above that the explanation that is provided by AI explainability methods is most often an account of what features in the input data are relevant for the classification task the AI system is trained to perform, and the weight the features have in the model. These explanations are general in the sense that they provide information about what input features are used when the system classifies something as X, or when it distinguishes between X and Y. They are not explanations of why the output was X, or X and not Y in a particular instance.

Here it is important to note a specific difference between AI systems and human thinking and clinical reasoning and decision-making. Most AI systems perform their classification tasks bottom up, i.e. from the data points in the input data to a classification. They are not contrastive. They do not explicitly compare different possible outcomes. The possible AI explanation of 'why X and not Y' will therefore differ in structure from the typical human explanation of the same question. The human diagnostic process may be seen as involving essentially an inference to the best possible explanation, i.e. an inference from a set of signs, symptoms and indicators to a conclusion taken to constitute the best possible explanation of these.(Dragulinescu, 2016; Lipton, 2017) Making an inference to the best possible explanation necessarily involves a consideration of whether there are alternative possible explanations that better explain the relevant set of signs, symptoms and indicators. Thus, diagnosis based on inference to best possible explanation specifically answers the question of 'why X and not Y'..

In the clinical context the AI explainability explanation will then in most cases be translated by the health care professional to an explanation the professional think is understandable by the patient[2]. This will in many cases involve a further simplification of the explanation. The explanation the patient receives will therefore often be doubly simplified. First by the AI explainability method employed, and then by the health care professional.

WHAT EXPLANATIONS DO PATIENTS WANT?

What types of explanations are patients likely to want in a diagnostic or treatment choice context when the health care professional has used an advisory AI system? Or to put it differently, what requests for explanation are they likely to raise?

Let us first note that because of the still existing general trust in health care professionals many and

2 We here assume that the professional has sufficient understanding of the explanation provided by the explainability method to be able to translate it and communicate it to the patient. This assumption is probably questionable in many instances but may become more realistic when more AI systems are implemented in health care and health care professionals become more acquainted with AI explainability information. There may also be a specific need to develop AI – Health care professional interfaces that support the understanding and translation of explainability information in the clinical context.(Amann et al., 2020; Holzinger & Müller, 2021)

perhaps even most patients will not put forward any requests for explanation but will accept the diagnosis and / or the treatment option put forward by the health care professional.(Nuffield Trust, 2022) However, some patients will require explanations..

Some of the likely requests for explanations are unrelated to the use of AI and may even lie outside the proper scope of the expertise of health care professionals. The professional will, for example have no proper answer or explanation to offer in response to the question 'Why have I got cancer?' when this question is raised in an existential mode.[3] If we concentrate on requests for explanation more closely linked to the diagnostic or therapy choice process and the involvement of AI in that process likely and very reasonable questions can include (questions here only enumerated in relation to diagnosis but similar questions could be asked in relation to therapy choice):

* Why do you think I have X?[4]
Does all the evidence point towards me having X?
What alternatives did you consider?
What did the AI advice?
* Why did you follow / not follow that advice?
How good is the AI?
Where does the AI get its information about me from, and what information does it use?
I have heard that some AI systems are racially biased, what do you know about this one?
Who has developed this AI, are there any conflicts of interest?
* I am surprised by the result, I thought it was much more likely to be Y? My neighbour has the same symptoms and her doctor told her it was Y.
* Can you provide an explanation of the AI advice?

There are also some relevant questions a patient could ask, but which require some knowledge about machine-learning, for instance:

What are the characteristics of the learning set of data that was used?
Has the AI performance been validated using a completely separate set of test data? Has it been validated against a data set derived from my/our population?

Some of these questions are straightforward requests for an explanation (marked with * above), but the type of explanation that is an appropriate response differ between them, and some may require both an explanation of the professional's thought processes and an explanation of the function of the AI in relation to the specific patient. Some of the other questions may, depending on how they are answered generate requests for explanations, e.g. the question about what information the AI uses may generate several distinct requests for explanation if the answer is 'your patient records, tax and credit card records and your social media data'.

What is the puzzlement or the epistemic interest that lies behind these requests for explanation to use Passmore's analysis? The range of possible epistemic interests is vast, but we will here focus on three distinct interests that are likely to be present when a patient is given a non-trivial diagnosis. One type of puzzlement or epistemic interest concerns the certainty of a diagnosis. The patient is about to make a significant medical decision and wants to have a high degree of epistemic certainty in relation to a major premise for this decision, i.e. the diagnosis. The patient wants to have a better understanding of what the epistemic warrant is for the claim that they suffer from X, or suffer from X and not Y. When AI advice has been involved in the diagnostic process and a patient raises a question about the certainty of the diagnosis, the answer will have to contain both an explanation of the accuracy of the advice, and an explanation of why the health care professional chose either to rely on the advice or not rely on the advice. A related type of puzzlement or epistemic interest could arise from surprise, i.e. the patient did not expect the diagnosis that is presented, but another diagnosis or a bill of clean health, or is surprised by the treatment options that are offered. The patient therefore asks for and need an explanation of why the situation is different from the one they expected. In this case the health care professional must again be able to provide an explanation of the accuracy of the AI advice and the role it played in their diagnostic decision-making, but must also be able to provide a 'contrastive explanation', i.e. an explanation of why the options or possibilities that the patient had in mind initially, before getting the diagnosis are not warranted by the evidence. Providing the contrastive explanation may be more difficult when AI is involved, because even AI systems that provide some information about why they classified the patient in a certain way, often do not provide any explanation of why other possibilities were discarded. A third type of puzzlement can be generated by suspicion and the belief that there is something wrong about the process, the way the outcome is generated, or the outcome itself. This suspicion can be pre-existing, or it can be generated by surprise in relation to the specific outcome. Here the health care professional must be able to explain why the process is robust and leads to reliable outcomes, but will often need to inquire further into the patient's specific concerns before providing this explanation. For example, a patient who is suspicious of commercial involvement in AI development has a different interest and needs a

3 Unless perhaps in the rare situation where professional and patient know each other well and belong to the same faith community.

4 This can be two different questions depending on whether the stress is on 'I' or on 'X' and would therefore call for two different explanations depending on which question is asked.

different explanation than the patient who is worried that health care professionals have not been properly trained to use AI systems in the right way..

In relation to all three types of puzzlement the kind of explanation that can be provided through explainable AI can often form part of the explanation, but as we have illustrated above more will also often be needed to adequately respond to all of the patient's epistemic interests e.g. accounts of the interaction between the AI and the professional or information about the AI system which is not about its inner workings but about how it was developed, what interests it potentially embeds, if and how it has been tested for bias across different groups etc.. The explainable AI explanation may also in itself not fully satisfy the patient's request for an explanation. Some patients will be satisfied by being told that a, b, c … z are the features the AI system uses to classify a patient as having X or needing treatment T, but some will want an explanation for why the system has classified them as having X or needing T. This request for an individual explanation may also be generated when a general explanation is provided and the patient notices that there seems to be a mismatch between that explanation and what they know about their own case, e.g. the explanation of the AI model states that feature f is important for classifying a patient as having X, but the patient knows that they have never had the investigation that provides the data for f. The honest health care professional will also have to acknowledge that the explanation of the AI advice they are providing and interpreting for the patient is a simplification and may not be an accurate reflection of the workings of the system.

In relation to the third type of puzzlement it is arguable that what the patient wants and needs is not merely an explanation but an effective ability to contest the outcome, the diagnosis or the treatment options they are presented with. The patient not only does not understand the outcome, they also think that it is wrong in some way, and they think that if it forms the basis for medical decisions they may be harmed or wronged. They thus have a strong interest in being able to effectively contest the outcome in order to protect against what they see as a potential threat to their health and welfare.(Ploug & Holm, 2020) Analysing the precise requirements of effective contestation of AI advice is beyond the scope of this paper, but they will go beyond explainability and may, in some cases involve a right to a second opinion. (Ploug & Holm, 2020, 2021, 2023)

CONCLUSION

In this paper we have argued that the types of explanation that is provided by current AI explainability methods do not adequately answer many reasonable requests for explanation that patients can make when their diagnosis or treatment choice has involved the use of AI advice. In relation to some requests these types of explanation are only part explanations, in other cases they answer the wrong question, and in yet other cases they are not explanations at all.

It is therefore important that health care professionals realise that patient queries in relation to AI involvement in diagnosis and treatment choice are not always, and perhaps even not in most cases requests for the kind of explanation that AI explainability methods can deliver. It is therefore the role of the health care professional when such queries are raised to inquire into and discern what it is that really puzzles the patient, and tailor the explanation that is provided to the patient's epistemic interests. The need to be able to provide a range of different explanations has implications for how we should train health care professionals to understand and use AI systems. Such training should not focus exclusively technical aspects, such as understanding how AI architectures work or what explainable AI can deliver, but should include the wider contexts of patients' epistemic interests and reasonable requests for relevant explanations.

For the patient who wants to contest an AI outcome or the professionals' adoption of that outcome as good advice their basis for contestation will only rarely be affected by the availability of an explanation generated by an explainability method. That basis is most commonly the belief that there is something questionable about the diagnosis or the treatment choice in the individual case, and the pieces of knowledge that have the potential to dispel that particular kind of puzzlement are often not about the internal workings of the AI system but about quite different things.

REFERENCES

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & the Precise4Q consortium. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

Bonkhoff, A. K., & Grefkes, C. (2022). Precision medicine in stroke: Towards personalized outcome predictions using artificial intelligence. *Brain*, *145*(2), 457–475. https://doi.org/10.1093/brain/awab439

Dragulinescu, S. (2016). Inference to the best explanation and mechanisms in medicine. *Theoretical Medicine and Bioethics*, *37*(3), 211–232. https://doi.org/10.1007/s11017-016-9365-9

Ferdous, M., Debnath, J., & Chakraborty, N. R. (2020). Machine Learning Algorithms in Healthcare: A Literature

Survey. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225642

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Holzinger, A., & Müller, H. (2021). Toward Human–AI Interfaces to Support Explainability and Causability in Medical AI. *Computer*, *54*(10), 78–86. Computer. https://doi.org/10.1109/MC.2021.3092610

Lipton, P. (2017). Inference to the Best Explanation. In *A Companion to the Philosophy of Science* (pp. 184–193). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405164481.ch29

Nirvik, P., & Kertai, M. D. (2022). Future of Perioperative Precision Medicine: Integration of Molecular Science, Dynamic Health Care Informatics, and Implementation of Predictive Pathways in Real Time. *Anesthesia & Analgesia*, *134*(5), 900. https://doi.org/10.1213/ANE.0000000000005966

Nuffield Trust. (2022). *Patient experience: Do patients have confidence and trust in clinicians?* Nuffield Trust. https://www.nuffieldtrust.org.uk/resource/confidence-and-trust-in-clinicians

Passmore, J. (1962). Explanation in Everyday Life, in Science, and in History. *History and Theory*, *2*(2), 105–123. https://doi.org/10.2307/2504458

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics—A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine*, *107*, 101901. https://doi.org/10.1016/j.artmed.2020.101901

Ploug, T., & Holm, S. (2021). Right to Contest AI Diagnostics: Defining Transparency and Explainability Requirements from a Patient's Perspective. In N. Lidströmer & H. Ashrafian (Eds.), *Artificial Intelligence in Medicine* (pp. 1–12). Springer International Publishing. https://doi.org/10.1007/978-3-030-58080-3_267-1

Ploug, T., & Holm, S. (2023). The right to a second opinion on Artificial Intelligence diagnosis—Remedying the inadequacy of a risk-based regulation. *Bioethics*, *37*(3), 303–311. https://doi.org/10.1111/bioe.13124

PricewaterhouseCoopers (PwC). (2018). *Explainable AI – Driving business value through greater understanding*.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, *28*(1), Article 1. https://doi.org/10.1038/s41591-021-01614-0

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv:1711.05225 [Cs, Stat]*. http://arxiv.org/abs/1711.05225

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (2016).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-Agnostic Interpretability of Machine Learning* (arXiv:1606.05386). arXiv. https://doi.org/10.48550/arXiv.1606.05386

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. *Proceedings of the 34th International Conference on Machine Learning*, 3145–3153. https://proceedings.mlr.press/v70/shrikumar17a.html

Subramanian, M., Wojtusciszyn, A., Favre, L., Boughorbel, S., Shan, J., Letaief, K. B., Pitteloud, N., & Chouchane, L. (2020). Precision medicine in the era of artificial intelligence: Implications in chronic disease management. *Journal of Translational Medicine*, *18*(1), 472. https://doi.org/10.1186/s12967-020-02658-5

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Hachette UK.

White House Office of Science and Technology. (2022). *Blueprint for an AI Bill of Rights – Making Automated Systems Work for the American People.* The White House. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

World Health Organisation (WHO). (2021). *Ethics and governance of artificial intelligence for health*. https://www.who.int/publications-detail-redirect/9789240029200

ACKNOWLEDGEMENTS

This research has been supported by a grant from The Independent Research Fund Denmark (10.46540/2027-00140B), a grant from The Poul Due Jensen/Grundfos Foundation and a grant from the Willum Foundation.

We also acknowledge Klitgården Refugium where the first draft of this paper was started.

COMPETING INTERESTS

The authors have no competing interest to declare.

Discussion Paper:

Does Living Ethically Make Life Meaningful? An Analysis from a Kantian Perspective

Hayate Shimizu
Department of Philosophy and Religious Studies, Hokkaido University

Abstract

This study explores how people should live their lives to make them meaningful, adopting a Kantian perspective. Specifically, this paper focuses on Kant's theory of virtue and argues that we can make our lives meaningful by living ethically. In the context of the philosophy of meaning in life, this view aligns with a form of objectivism, which holds that life becomes meaningful through the achievement of moral value in accordance with objective norms. Additionally, this paper incorporates elements of subjectivism into the objectivist framework, suggesting that living an ethical life makes life meaningful, emphasizing that subjective satisfaction derived from following moral norms is combined with a meaningful life. Drawing on Kant's argument that a virtuous agent experiences a sense of self-satisfaction as a reward for leading an ethical life, this approach is termed the Kantian hybrid theory. Finally, this paper interprets life as meaningful when we are satisfied with our state of being, which results from living ethically.

Keywords: Ethics, Meaningful life, Subjectivism and objectivism, Kant, Virtue

1. Introduction

In recent years, the concept of meaning in life has primarily been explored through the lens of analytic philosophy. Notably, Metz's *Meaning in Life* (2013) has led several philosophical investigations into how we understand and make sense of meaning in our lives. Accordingly, *The Oxford Handbook of Meaning in Life* (2022) shows an increase in research into this topic from various approaches. Despite its significance as a philosophical issue, however, relatively few studies have examined meaning in life from historical or classical philosophical perspectives.

This study explores how people should live their lives to make them meaningful, adopting the classical perspective of Kantian ethics while drawing on the existing philosophical framework of meaning in life. This paper aims to argue, based on Kant, that the life of a virtuous person who lives ethically is meaningful. It explores the resources that Kant's theory provides for the philosophy of meaning in life. Accordingly, it only briefly engages with the framework of meaning in life and instead focuses on Kant's theory of virtue. Kant's theory of virtue can provide a theoretical framework for answering the question: 'Does living ethically make life meaningful or meaningless?' This question is crucial to the philosophy of meaning in life. Here, we interpret 'living ethically' as ' following moral norms, even at the expense of one's desires.' The abovementioned question can be paraphrased as follows: 'Is life made meaningful by doing what we ought to do objectively, even at the expense of our subjective satisfaction?' Kant's answer would be yes. This question is rooted in objectivism, following Metz's distinction between 'subjectivism' and 'objectivism.' How exactly would Kant fit into Metz's categories of philosophy of the meaning in life? This study aims to show that Kant's argument based on his theory of virtue can address subjective satisfaction

that makes life meaningful while following the line of objectivism.

This paper is organized as follows. Section 2 briefly reviews subjectivism and objectivism, discussing their respective strengths and limitations. Section 3 reviews several studies on the meaning of life from the perspective of Kant's philosophy and provides an outline of Kant's theory of virtue. Section 4 focuses on Kant's argument that virtuous agents attain a sense of satisfaction with themselves as a reward for consistently living a virtuous, ethical life. This sense of satisfaction, which Kant describes as *moral satisfaction*, arises from living ethically. This satisfaction is experienced subjectively; however, unlike mere sensible pleasure, a commitment to objective morality is essential to obtain this satisfaction, as living ethically is a condition for moral satisfaction. We call this position the *Kantian hybrid theory*, which is a framework that integrates a subjectivist element within the broader structure of objectivism.

2. Subjectivism and objectivism

In a common classification, the philosophy of meaning in life is divided into two main categories: subjectivism and objectivism. Broadly speaking, these views differ in whether the meaning of life is determined by subjective or objective factors. Subjectivism holds that a life is meaningful if the person living it feels satisfied, even if their life does not fulfill any objective standards of greatness. In contrast, objectivism holds that a life is meaningful if it is objectively great, regardless of whether the person feels satisfied. Additionally, a third perspective, known as the hybrid theory, suggests that both subjective and objective factors must be combined to determine life's meaning. Wolf, a leading proponent of this view, has contributed significantly to this strand of research. We will now examine the key characteristics of subjectivism and objectivism.

According to subjectivism, meaning in life depends on each individual's variable pro-attitudes.(1) In other words, subjective satisfaction is the only factor determining meaning in life. From this perspective, meaning in life is 'mind-dependent,' as it is determined by individuals' positive attitude. The central argument for this perspective is Taylor's (1970) discussion of the myth of Sisyphus. According to Taylor, a life spent merely rolling a rock up a hill may appear to us as soulless and lacking any impact on the world; however, if Sisyphus rolls the rock up the hill and finds joy and satisfaction in it, then his life is meaningful.(2) In other words, only satisfaction, which depends on the individual's mind, determines the meaningfulness of life. However, this idea highlights the following shortcomings of subjectivism. If the meaning in life is determined only by the satisfaction that depends on the individual's mind, then an evil person like Hitler could be said to have had a more meaningful life than someone who devoted their life to serving others, which involves continuous struggle.(3) This conclusion seems counterintuitive, and it is difficult to conceive that the meaningfulness of life is contingent, depending on the state of the subject.

In contrast, according to objectivism, meaning in life is not determined by subjective factors, but by the extent to which a person has achieved something of objective value.(4) Objective values refer to truth, goodness, and beauty.(5) In other words, the determinants of meaning in life are independent of individual satisfaction. Therefore, in contrast to subjectivism, objectivism conceives of the meaning in life as 'mind-independent.' While what is objectively valuable may not be singularly determined, this paper focuses primarily on moral value. From the perspective of objectivism, which focuses on achieving the moral value, a meaningful life should be defined as a life dedicated to making society a better place by doing what is ethically good according to objective moral norms.(6) Metz provides the example of Mother Teresa's life as an illustration of this principle.

According to this view, our lives become meaningful when we achieve the objective value of morality, even at the expense of our subjective pleasures. This makes accepting objectivism particularly challenging, as it can seem overly demanding. It may seem excessively harsh to claim that a life is meaningless if one does not live ethically. If only an ethical life has objective value, the life of someone who has failed to live ethically may be deemed devoid of value or meaning. However, can an ethical life be approved as meaningful to a person at the expense of subjective satisfaction? Is not a certain amount of satisfaction and fulfillment necessary to make life more meaningful? As Metz (2013) and Kauppinen (2012) point out, when we talk about the meaning in life, we necessarily connote an emotional response. Kauppinen refers to it as 'feelings of fulfillment and

(1) cf. Metz 2022. Note that 'subject' here means a way of being subject that allows one to take a 'propositional attitude.'

(2) cf. Taylor 1970, 323.

(3) Metz 2013, 175-176.

(4) It is not necessary to assume that subjective factors do not influence meaning in life at all. Metz, for example, discusses meaning in life from an objectivist perspective but considers that the addition of subjective satisfactions further enhances the meaning.

(5) cf. Metz 2022.

(6) While it may be anachronistic to fit it into the framework of objectivism, I believe that even proponents like Singer, who argue that living ethically is the best way to achieve a meaningful life, would find resonance with this idea (Singer 1993).

admiration being appropriate' (Kauppinen 2012, 353). The sense of fulfillment that comes from doing what is ethically good is precisely the emotional response that justifies life's meaningfulness.

This study argues that living ethically makes life meaningful not only because it has an objective (moral) value in itself but also because it involves the sense of fulfillment and satisfaction with the self. To this end, this study adopts a classical Kantian perspective to show that an ethical or virtuous life is meaningful. It is the perspective of Kantian ethics that can provide the answer that a sense of satisfaction or fulfillment with the self is an important emotional response to the meaning in life, and that this is obtained by living ethically. This study proposes a Kantian hybrid theory that justifies, within the framework of objectivism, the idea that subjective satisfaction is an important element in making life meaningful.

3. The problem of the meaning of life and virtue in Kant

3.1. Did Kant take issue with the meaning of life?(7)

Several works discuss Kant's philosophy and the meaning of life. For example, Godlove's (2018) 'Kant and the Meaning of Life' focuses on the argument of the highest good being the ultimate end of action and takes God as a requirement of practical reason, as developed in Kant's *Critique of Practical Reason* (1788). For an agent's life to be meaningful, they must promote a world in which happiness is apportioned according to virtue (cf. 05, 110)(8). In Kant's moral philosophy, the concept of the highest good includes the idea that happiness must be distributed in proportion to virtue. Therefore, for Kant, the meaning of life is the pursuit of the highest good(9). However, since this distribution cannot be guaranteed, Kant argues that the existence of God must be postulated. God, as the moral distributor, ensures that happiness is in proportion to virtue, thereby fulfilling the requirements of the highest good. Hence, he concludes that the meaning of life is not possible without the existence of God.

Some works have discussed the meaning of life in Kantian philosophy differently, such as Church's (2022) *Kant, Liberalism, and the Meaning of Life*. Part 1 analyzes Kant's view of the meaning of life, centering on the (moral) progress of humanity as a whole rather than individual lives. As he states, 'For Kant, our lives gain meaning through participating in and contributing to relationships and institutions which aim to advance humanity's progress.' (10) He considers the meaning of life in terms of the moral purpose of contributing to the community, humanity, or the world. According to Kant, life is meaningful because it contributes to the moralization of humanity and the building of the Kingdom of Ends.

Based on Metz's categories, Kant's view of the meaning of life can be seen as a form of objectivism. However, Kant acknowledges the role of subjective satisfaction in life through ethical living, which is a subjectivist element. Therefore, this paper argues that Kantian virtuous agents make their lives meaningful by gaining subjective satisfaction through an ethical life guided by objective morality. This allows us to address the question of how subjective factors are involved in meaning in life from a Kantian perspective. This theory offers a comprehensive framework by integrating objectivism and subjectivism, both of which can be justified from a Kantian perspective..

Although categorizing Kant within the frameworks of objectivism and subjectivism in the context of the meaning of life is anachronistic, it is possible to reconstruct Kant's perspective on this topic. To this end, we will focus on Kant's theory of virtue. As Kant believes that virtuous agents are cheerful and satisfied with themselves, this argument can be connected to a discussion on meaning in life. This connection justifies the claim that being virtuous makes life meaningful, objectively valuable ('achievement of the moral value'), and subjectively valuable ('satisfaction with oneself'). In this way, the compatibility of objectivism and subjectivism can be explored. With this perspective in mind, we will review Kant's conception of virtue.

3.2. Kantian Virtue

Kant discusses virtue mainly in his *Doctrine of Virtue*, stating that a defining characteristic of virtue is the 'strength' of will required to fulfill one's duty. Kant defines virtue as follows: 'Virtue is the strength of a human being's maxims in fulfilling his duty' (06, 394) and 'Virtue is, therefore, the moral strength of a human being's will in fulfilling his duty' (06, 405). Kant characterizes virtue as a strength because it requires an individual to maintain a moral disposition based on duty, despite the temptations of emotions and inclinations. In

(7) Since the phrase used by the previous studies to which this paper refers is 'meaning of life,' this section adapts it.)

(8) Quotations from Kant's works cite the volume and page number of the Academy edition, Kants Gesammelte Schriften, ed. Königlich Preussische Akademie der Wissenschaften, vols. 1–29, Berlin: de Gruyter, 1902.

(9) Godlove 2018, 147.

(10) Church 2022, 98.

other words, virtue involves sacrificing one's subjective pleasure to fulfill objective duties.

It follows that virtue is a 'moral disposition in conflict' between duty and the inclination to rebel against it.(11) As Kant explains, 'his [i.e., human beings'] proper moral condition, in which he can always be, is virtue; that is, moral disposition in conflict, and not holiness in the supposed possession of a complete purity of dispositions of the will' (05, 84). Virtue must take the form of conflict because humans are not pure and are affected by inclinations that oppose moral laws. In other words, Kantian virtue presupposes the strength of will, as determined by moral law, and therefore necessarily involves limiting one's feelings. It requires governing one's subjective feelings through reason and objective moral norms. Therefore, from a Kantian perspective, an agent following a virtuous life obeys objective ethical rules via a strong will. In this sense, Kant's virtuous agents do not seem to fulfill the subjectivist meaning in life. At the very least, the happiness from subjective and emotional satisfaction with one's condition is not fulfilled in the moral life.

Moreover, according to Kant's critique of eudaimonism, satisfaction based on subjective feelings seems to be incompatible with a virtuous and ethical life.(12) Does the virtuous Kantian agent lead to a painful and unpleasant life at the expense of satisfaction and pleasure? Are they wholly denied the meaning in life, as subjectivism claims? However, subjective elements such as pleasure and a sense of satisfaction are not excluded from Kant's theory of virtue. As scholars like Sherman (1997) have noted, Kant did not entirely exclude emotions from the realm of morality. Kant's criticism of eudaimonism rejects the notion that happiness, understood as the satisfaction of one's inclinations, can serve as a moral principle(13). However, this does not imply a complete rejection of the emotional states associated with virtue. As we will explore in the next section, some scholars have argued that the Kantian moral life involves moral pleasures such as self-satisfaction. This paper examines this argument in the context of the philosophy of meaning in life and evaluates it from the perspectives of objectivism and subjectivism. In this context, an approach based on Kant's theory of virtue can be attractive in that it defends the meaning in life in a broad objectivist framework while preserving subjective satisfaction. This is because, while respect for objective moral values in the Kantian sense is a necessary condition for a meaningful life, the subject also obtains subjective satisfaction. We call this the Kantian hybrid theory. In the next section, we defend the position that a virtuous or ethical way of life makes a person's life meaningful.

4. Kantian Virtuous Agents Lead Meaningful Lives

The life of Kant's virtuous agent must be committed to objective moral values. Further, such agents must have a strong will, that is, a tranquil mind [*Gemüth in Ruhe*] that maintains that good is determined by moral law. However, this does not necessarily mean that they forgo all subjective satisfaction. In the *Critique of Practical Reason*, Kant had already indicated that the feelings associated with virtue are self-contentment. Kant says, 'I certainly do not deny that frequent practice in conformity with this determining ground can finally produce subjectively a feeling of contentment with oneself [*Zufriedenheit mit sich selbst*]' (CPrR, 05: 38). In addition, Kant suggests the presence of emotional states as analogs of happiness, distinct from happiness and necessarily associated with virtue, which he describes using the term 'satisfaction with oneself':

> Have we not, however, a word that does not denote enjoyment, as the word happiness does, but that nevertheless indicates a satisfaction with one's existence, an analogue of happiness that must necessarily accompany consciousness of virtue? Yes! This word is satisfaction with oneself, which in its strict meaning always designates only a negative satisfaction with one's existence, in which one is conscious of needing nothing. Freedom, and the consciousness of freedom as an ability to follow the moral law with an unyielding disposition, is independence from the inclinations, at least as motives determining (even if not as affecting) our desire, and so far as I am conscious of this freedom in following my moral maxims, it

(11) In this battle, the enemy opposed to morality is inclination. However, it is important to note that the battle is not about overthrowing inclination itself because it is 'bad.' As Kant puts it, 'considered in themselves, natural inclinations are good' (06, 58).

(12) In the *Critique of Practical Reason*, for example, Kant assumes that happiness is contingent and dependent on personal feelings and desires, which is necessarily incompatible with the objectivity of moral values. Kant discusses the critique of eudaimonism in the context of the ancient concept of happiness (eudaimonism). See Irwin's (1996) 'Kant's Criticisms of Eudaemonism' for more information.

(13) Kant's concept of happiness in his critique of eudaimonism is limited. For example, as Elizondo (2023) points out, we can see that, while Aristotle is not a eudaimonist by Kant's lights, Kant is a eudaimonist by Aristotle's lights. A separate discussion is needed to explore in what sense Kant critiques eudaimonism, but that issue lies beyond the scope of this paper.

> is the sole source of an unchangeable satisfaction, necessarily combined with it and resting on no special feeling, and this can be called intellectual satisfaction. (05, 117-118)

Satisfaction with oneself does not mean the pleasure that comes from the fulfillment of one's feelings and desires. Instead, it is what one can feel toward oneself with the consciousness that one has achieved freedom in the Kantian sense of observing the moral law independently of one's desires (inclination). Hence, this satisfaction with oneself is not based on emotion but accompanies virtue. Nevertheless, the virtuous person has pleasure in the moral sense. Kant further adopted this view in his *Doctrine of Virtue*, where he refers to 'a subjective principle of ethical reward: […] that is, a receptivity to being rewarded in accordance with laws of virtue: the reward, namely, of a moral pleasure that goes beyond mere satisfaction with oneself (Zufriedenheit mit sich selbst) (which can be merely negative) and which is celebrated in the saying that, through consciousness of this pleasure, virtue is its own reward' (06, 391). This argument aligns with the idea that self-satisfaction necessarily accompanies the consciousness of virtue. According to Kant, the state of contentment can only be realized as a reward when it follows the awareness of having fulfilled one's duty through the exercise of reason. The feeling of moral pleasure as a reward occurs only as a result of virtue, meaning it always follows virtue rather than preceding it (cf. Cohen 2018, 15; Walschots 2017).

According to Kant, the virtuous agent who aims to overcome opposition in the conflict of virtue and fulfil his duty without the influence of emotion is in 'a state that could well be called happiness, a state of satisfaction and peace of soul in which virtue is its own reward' (cf. 06, 377). Satisfaction, which is inevitably linked to a virtuous life, can be interpreted as making life more meaningful. Kant does not explicitly refer to this satisfaction as the meaning of life. However, in the context of the philosophy of the meaning in life, we may interpret the state of self-satisfaction derived from a virtuous life as constituting meaning in life. Therefore, we can find our lives meaningful as a reward for living an ethical life.

Moreover, in Section II of the *Doctrine of Virtue*, Kant introduces a chapter on 'ethical ascetics,' where he asserts that one of the frames of mind to be aimed for in the cultivation of virtue is a cheerful frame of mind in fulfilling one's duties. According to Kant's theory of virtue, a virtuous agent is not in a gloomy mood and is not excluded from the joys of life. Instead, the life of a virtuous person must be cheerful, as an agent without pleasure does not have a joyous heart and, as such, is not virtuous. Kant states that 'a heart joyous in compliance with its duty is the sign of the genuineness in a virtuous disposition' (06, 24). However, this joyous state of mind differs from mere sensible satisfaction. This state of self results from autonomous choice based on reason. In other words, the positive emotional state that results from achieving freedom based on reason characterizes a virtuous person. Therefore, virtue training should aim to achieve this state. Kant states:

> The rules for practicing virtue (exercitiorum virtutis) aim at a frame of mind that is both valiant and cheerful in fulfilling its duties (animus strenuus et hilaris). For, virtue not only has to muster all its forces to overcome the obstacles it must contend with; it also involves sacrificing many of the joys of life, the loss of which can sometimes make one's mind gloomy and sullen. (06, 484)

Why should a virtuous person be cheerful? The reason isto avoid moodiness while observing one's duties. If a virtuous life were entirely painful, it would not be of any value to people and would be a life that everyone avoids. Therefore, Kant believes that a virtuous life must be meaningful and have a positive value. Kant states:

> […] if duty is not done with pleasure (mit Lust) but merely as compulsory service (bloß als Frohndienst), [it] has no inner worth for one who attends to his duty in this way and such service is not loved by him; instead, he shirks as much as possible occasions for practicing virtue. (06, 484)

Human virtue refers to a conflict between reason and emotion. Nevertheless, in the course of this conflict, one can become aware that one has overcome sensible impulses, which produce cheerfulness. As Kant puts it, 'Hence it makes one valiant and cheerful in the consciousness of one's restored freedom' (06, 485). Thus, the consciousness of freedom that arises from overcoming opposition in the conflict of virtue and establishing the self-governance of reason gives the virtuous agent a sense of satisfaction that is independent of inclinations. Virtuous agents accomplish virtue through reason, thereby achieving cheerfulness. This results in satisfaction with one's state of mind, even though it binds one's desires. This satisfaction is associated with the meaningfulness of life.

What is the source of the positive feelings of satisfaction and cheerfulness derived from being virtuous? This is undoubtedly distinct from the pleasures assumed by naïve subjectivism. A virtuous agent suppresses sensible emotions and desires in accordance with universal obligation. However, while the virtuous agent is limited in pleasure in this sense, it is somewhat

'elevated' by having a will determined by moral law. This concept is based on the dualistic view of humans presupposed by Kant. In other words, as a sensible being, one can feel pain, while as a rational being, one adoptse a positive attitude towards self-affirmation.(14) Therefore, the Kantian virtuous agent, while limited in pleasure in the sensible sense, is full of self-affirmation as a rational being.

To sum up, Kant's virtuous agents choose to live ethically by committing themselves to the moral law. This life may be unsatisfactory because it is limited to many subjective pleasures. However, virtuous agents experience moral pleasures that are distinct from sensible pleasures and, in this sense, they feel satisfied. In a special sense, this pleasure is an ethical reward that only a virtuous person can receive from being virtuous—i.e., living ethically. This reward can be reconstructed as one component of making life meaningful. The condition for life to be meaningful is to have the ethical resolve to continue committing to an objective moral norm using reason. While this may not be a sensuously pleasurable experience, it provides a sense of self-fulfillment. Living ethically limits some subjective pleasures and conditions commitment to objective moral values, while simultaneously, it gives the person a sense of moral satisfaction and makes life meaningful. Therefore, Kant's theory of virtue can be construed as a hybrid theory, adding an element of subjectivism to the objectivist line.

For Kant, living ethically is right, good, and meaningful. From this Kantian perspective, two key components make life meaningful: (1) having a strong will to fulfill objective duties (objectivism) and (2) achieving a sense of satisfaction in one's state of mind (pro-attitude; subjectivism). This Kantian hybrid theory of meaning in life can serve as a theoretical basis for the claim that 'ethical life is meaningful.'

(14) 'As submission to a law, that is, as a command (indicating constraint for the sensibly affected subject), it therefore contains in it no pleasure but instead, so far, displeasure in the action. On the other hand, however, since this constraint is exercised only by the lawgiving of his own reason, it also contains something elevating [Erhebung], and the subjective effect on feeling, inasmuch as pure practical reason is the sole cause of it can thus be called self-approbation [Selbstbilligung] with reference to pure practical reason, inasmuch as he cognized himself as determined to it solely by the law and without any interest, and now becomes conscious of an altogether different interest subjectively produced by the law, which is purely practical and free' (05, 80-81). This could be reconstructed from the feeling of 'respect,' which Kant discusses in *The Critique of Practical Reason*. However, since this section aims to examine why virtuous agents have a pro-attitude, we will leave that argument aside.

5. Conclusion

Kant did not explicitly claim that only a virtuous life is meaningful. However, by applying Kant's theory of virtue to the philosophy of meaning in life, the Kantian hybrid theory proposed in this study argues that we can make our lives meaningful by living ethically. Kantian hybrid theory presupposes a commitment to the position of objectivism in a broad sense, in that it holds that a good and meaningful life is one in which one continues to fulfill acts that involve objective moral values, while simultaneously asserting that such a life is accompanied by subjective satisfaction. Therefore, this study posits that for life to be meaningful, it must be accompanied by subjective satisfaction, which is attained through ethical living. The Kantian hybrid theory supports the interpretation that meaning in life is a combination of objective moral values and subjective moral satisfaction. In this sense, a virtuous Kantian agent is able to make their life more meaningful because one of the best ways to make life meaningful is to live ethically.

However, this paper leaves a few issues unresolved. While this study concludes that living ethically, according to Kant's theory of virtue, makes life meaningful, this is only a Kantian response. It does not fully address the implications of this conclusion for the broader philosophical discourse on the meaning in life currently under discussion. In addition, the connection between a good life, happiness, and the meaningfulness of life has been a topic of discussion in eudaimonistic ethics since ancient Greece, with thinkers such as Epicurus, the Stoics, and Aristotle. Nevertheless, despite disagreements regarding the scope of ethical life, some thinkers believe that this is the path to achieving a meaningful life. At the very least, the conclusion of this paper provides a critical perspective when considering the issues of meaning in life and ethics.

References

Elisondo, E. S. (2023). Kantian Eudaimonism. Journal of the American Philosophical Association, 9(4), 655–669.

Church, J. (2022), *Kant, Liberalism, and the Meaning of Life*. Oxford: Oxford University Press.

Cohen, A. (2018), Kant on moral feelings, moral desires and the cultivation of virtue. In S. Sedgwick & D. Emundts (Eds.), *Begehren / Desire* (pp. 3–18). De Gruyter.

Irwin, T. H. (1996), Kant's Criticisms of Eudaimonism. In Engstrom, Stephen and Whiting, Jennifer (eds.), *Aristotle, Kant, and the Stoics: Rethinking Happiness and Duty*. Cambridge: Cambridge University Press, 63–101.

Kant, I. (1900ff), *Gesammelte Schriften* Hrsg.: Bd. 1–22 Preussische Akademie der Wissenschaften, Bd. 23 Deutsche Akademie der Wissenschaften zu Berlin, ab Bd. 24 Akademie der Wissenschaften zu Göttingen. Berlin.

(Translations of Kant's works are from the *Cambridge Edition of the Works of Immanuel Kant* (Cambridge University Press)).

Kauppinen, A. (2011), Meaningfulness and Time. *Philosophy and Phenomenological Research* 84 (2):345-377.

Landau, I (ed.). (2022), *The Oxford Handbook of Meaning in Life*. New York: Oxford University Press.

Godlove, T. F. (2017), Kant and the Meaning of Life. In Leach, Stephen and Tartaglia, James (eds.), *The Meaning of Life and the Great Philosophers*. London: Routledge, 142-149.

Metz, T. (2013), *Meaning in Life: An Analytic Study*. New York, NY: Oxford University Press.

—— (2022), The Meaning of Life, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/win2022/entries/life meaning/>.

Sherman, N. (1997). *Making a Necessity of Virtue: Aristotle and Kant on Virtue*. Cambridge: Cambridge University Press.

Singer, P. (1993), *How Are We to Live?: Ethics in an Age of Self-Interest*. Amherst, N.Y.: Oxford University Press.

Taylor, R. (1970), The Meaning of Life, in *Good and Evil*, repr. Amherst, NY: Prometheus Boooks, 2000: 319-34.

Walschots, M. (2017), Kant on Moral Satisfaction. *Kantian Review* 22 (2):281-303.

Notes to Contributors

1. All submitted papers are subject to anonymous peer-review, and will be evaluated on the basis of their originality, quality of scholarship and contribution to advancing the understanding of applied ethics and philosophy.

2. Papers should be at most 8,000 words, including references.

3. An abstract of 150-300 words and a list of up to 5 keywords should be included at the beginning of the paper.

4. Papers can be submitted any time of the year by e-mail to jaep@let.hokudai.ac.jp . If the authors wish their papers to be included in the next volume (to be published in March 2026), however, they are advised to submit their papers by September 15th 2025.

5. In-text references should be cited in standard author-date form: (Walzer 1977; Kutz 2004), including specific page numbers after a direct quotation, (Walzer 1977, 23-6).

6. A complete alphabetical list of references cited should be included at the end of the paper in the following style:

   Cohen, G.A. (1989), 'On the Currency of Egalitarian Justice', Ethics, 99 (4): 906-44.
   Kutz, C. (2004), 'Chapter 14: Responsibility', in J. Coleman and S. Shapiro (eds.), *Jurisprudence and Philosophy of Law*, Oxford, UK: Oxford University Press, 548-87.
   Walzer, M. (1977), *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, New York: Basic Book.

7. Accepted papers will be published in both web-based electronic and print formats.

8. The Editorial Board reserves the right to make a final decision on publication.