SAPERE AUDE

An undergraduate philosophy journal

2022 Issue

SAPERE AUDE

dare to know

‘*Sapere Aude*’, as used in Immanuel Kant’s essay, ‘What is Enlightenment?” means ‘Dare to Know.’ This phrase exemplifies the mission of *Sapere Aude*. Our aim is to facilitate intellectual discovery by encouraging undergraduate students to reason independently and to explore unfamiliar philosophical territory.

We invite undergraduate students to submit papers in all areas of philosophy annually. The papers should exhibit independent, creative thought and exemplify deep understanding of a philosophical subject. Submissions with interdisciplinary engagements are also encouraged (e.g. philosophical intersections with social sciences, humanities, natural sciences, social justice studies, etc.) Content, however, should be concerned with a primarily philosophical issue.

Sapere Aude begins taking submissions in September, and stops taking submissions around February.

Email sapere_aude@wooster.edu with any inquiries. Updated information on the status of our annual submission cycle or access to prior publications can be found via Facebook at:

https://www.facebook.com/SapereAudeCOW/

Or via our website:

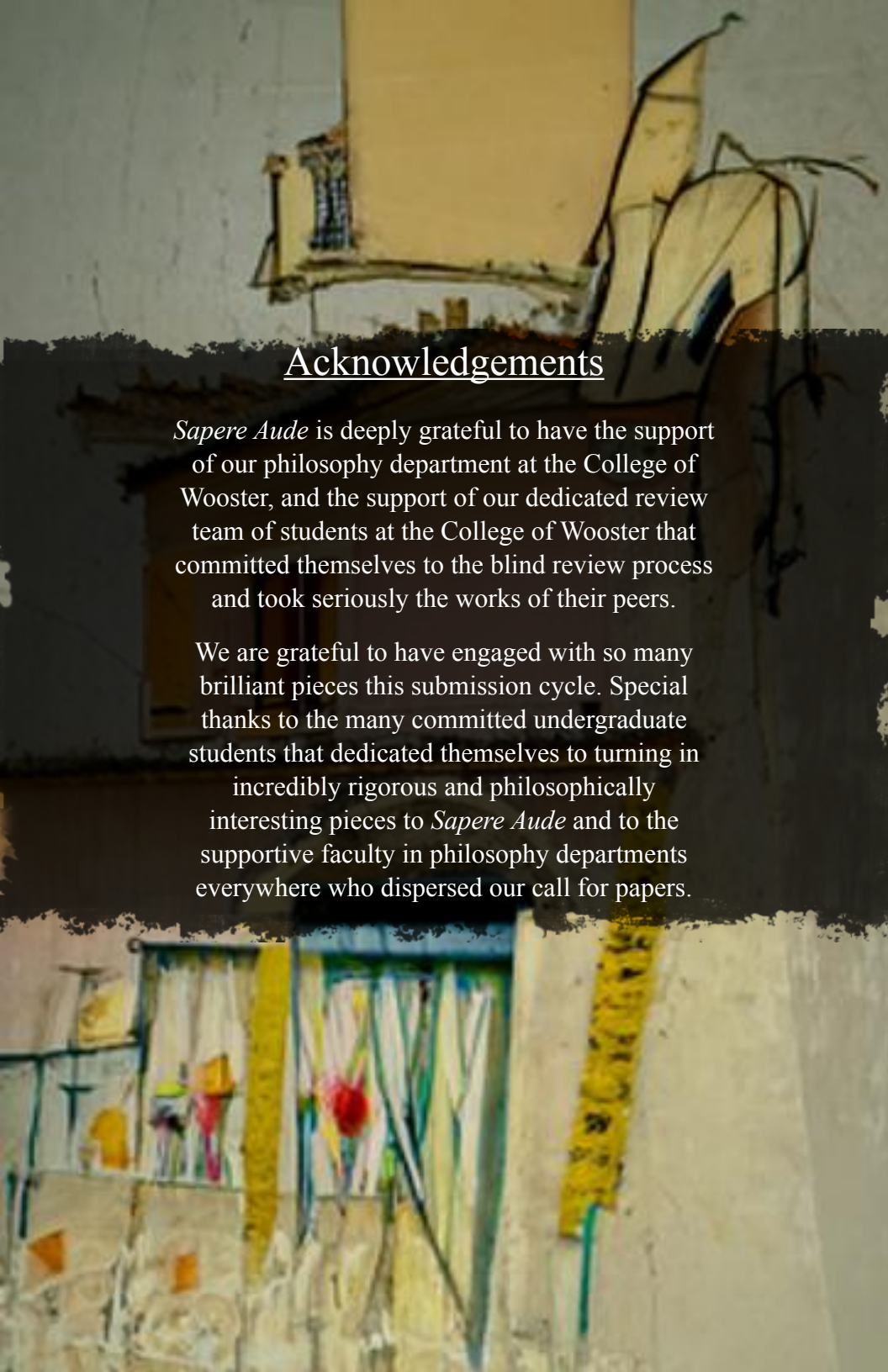
https://sapereaude.voices.wooster.edu

AI GENERATED ART

The 2022 issue of *Sapere Aude* features a cover as well as all article illustrations generated by NightCafe and Dall-e, AI art platforms that generate images based off of keywords via algorithms that analyze images as data points and respond to prompts with keywords according to the images attached to any given keyword. The prompt utilized for this cover was simply - *Sapere Aude*. Each article's related illustration was generated from the keywords in the titles themselves.

The emergence of AI generated artwork and its widespread accessibility has been widely controversial and discussed within philosophy at length. The supposed removal of the artist from the artwork and absolute algorithmization of the creation of art has raised many interdisciplinary philosophical questions - is artwork without an artist still art? What is the aesthetic value of AI art itself? Who truly owns AI generated art, especially when it draws upon the style or essence of other traditional artists? Is AI art exploiting the labor and creativity of artists by treating their intellectual property as data points?

The controversial nature of AI generated art itself made it a fitting medium for this year's edition of *Sapere Aude*, by exemplifying interdisciplinary philosophical inquiry that is grounded in the discourse of much of the community in 2022.



Acknowledgements

Sapere Aude is deeply grateful to have the support of our philosophy department at the College of Wooster, and the support of our dedicated review team of students at the College of Wooster that committed themselves to the blind review process and took seriously the works of their peers.

We are grateful to have engaged with so many brilliant pieces this submission cycle. Special thanks to the many committed undergraduate students that dedicated themselves to turning in incredibly rigorous and philosophically interesting pieces to *Sapere Aude* and to the supportive faculty in philosophy departments everywhere who dispersed our call for papers.

STAFF

Editor-In-Chief

Savannah Sima

Associate Editors

Artemis Swanson

Peter Barker

Content Editors

Peter Barker

Savannah Sima

Maxwell Hossler

Assistant Editorial Board

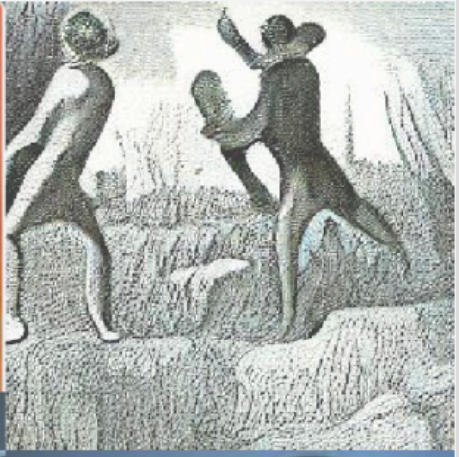
Aidan L'hommedieu

Artemis Swanson

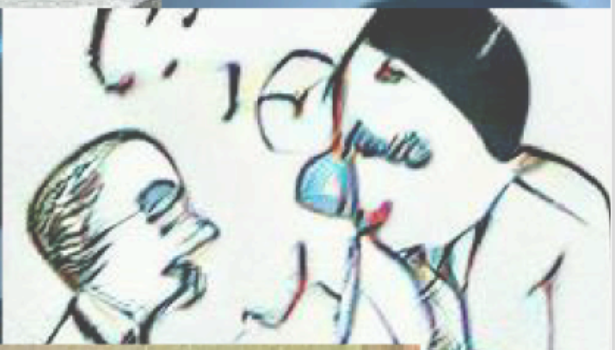
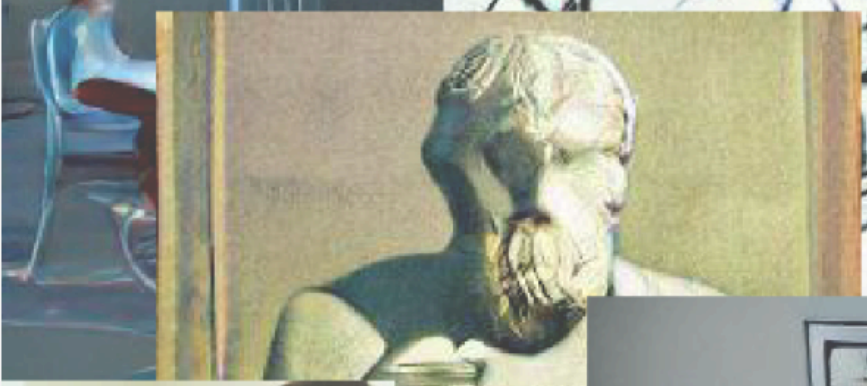
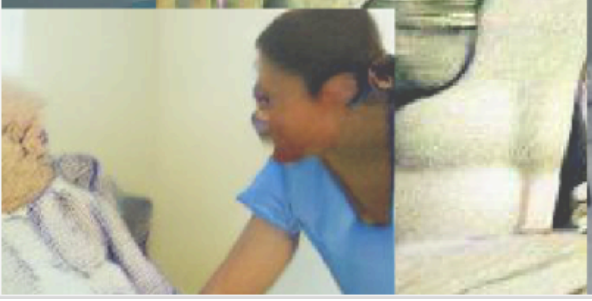
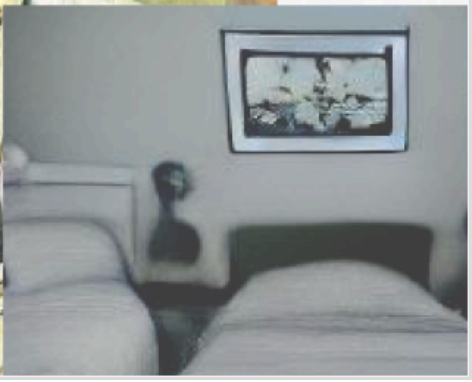
Ben Read

Jules Gerdes

Peter Barker



Whither is God?
Nietzsche's Madman and Ideology

Russell Clarke

Abstract

In his parable of *the Madman,* Nietzsche proclaimed the Death of God and, as such, symbolized the rejection of Christianity as the prevailing moral foundation of society during this period. This paper is an attempt to trace a genealogy of Nietzsche's thought on the topic of morality and nihilism in the face of God's symbolic death. I input the insights of Carl Jung, Louis Althusser, and Slavoj Žižek with respect to ideology and its effect on the subject. Moreover, I analyze the ways in which the object of ideology and its possession of the subject has changed the subject. This paper is a meditation on the question of ideology and its relation to individual agency. Ultimately, I wish to entertain the question of whether Nietzsche was right to question humanity's ability to cultivate their own will to truth.

You can take away a man's god,
but only to give him others in return
— Carl Jung, *The Undiscovered Self,* 1958

I. Introduction

In his work of philosophy Friedrich Nietzsche makes numerous attempts at diagnosing human proclivities through tracing and dissecting Truth-value systems with their relation to action, survival and fitness. While Nietzsche held derisive attitudes towards foundational truths as such, his works are also intrinsically epistemological and literary in their ability to utilize metaphor as a tool for the derivation of certain axioms which scaffold the Truth-value systems wherein lies the seemingly ineradicable ideological and religious structures which the Western world has held dear for over two millennia. The focal point, however, to Nietzsche's work, and more specifically his parable of *The Madman* in section 125 of *The Gay Science*, is religion and God as structural, moral and psychological embodiments which simultaneously delineate moral axioms and regulate psychosocial locomotion and hierarchy.

In this paper, I analyze Friedrich Nietzsche parable of *The Madman* and assess Nietzsche's attitudes toward the relationship of religion and social morality before and after the declaration of God's death. In doing so, I make the claim that in the wake of God's death and the concomitant erosion of foundational religious principles for new secularized ideological principles, the individual has undergone depotentiation by the process of ideological deification. I integrate Jungian insights on the topic of ideology and the repression of the *shadow* to demonstrate how *Nietzsche* was correct in his assumptions of society's passive nihilism. This paper also offers a variegated interpretation of ideological deification by utilizing structural and psychoanalytic accounts of ideology and the subject's relation to it. I then tie these components together with a generous discussion of contemporary social and political ideologies and the individual's capacity, or incapacity, to extricate himself from ideology.

II. The Madman's Declaration

Nietzsche's attitude toward God (the deity himself) is complex, to say the least. I say "the deity himself" to distinguish between his attitudes towards the metaphysical possibilities of a deity as opposed to his attitudes towards the herd morality and religious psychology that underpin 'God' as an idea. What we do know, however, is that Nietzsche abhorred Christianity and dedicated much of his writings to burlesquing it, which reaches a climax when in the parable, an eccentric man, standing before an audience holding a lantern aloft, asks 'Whither is God?"[1] This is implored rhetorically as the conclusion is already reached by the madman that "*we have killed him*—you and I."[2] The man is regarded as mad precisely because of his irrational preoccupation with God's death, a matter of indifference to those in the marketplace to whom the declaration is made and whose return to this declaration is laughter and derision.

Nietzsche uses imagery and metaphor to illustrate the self-inflicted condition of the society he is analogizing where the traditional religious God-concept has been subverted and negated. "Who gives us the sponge to wipe away the entire horizon?"[3] the madman asks. "What were we doing when we unchained this earth from its sun?"[4] he implores. In likening the death of God to the erasure of a horizon or the suspension of the Sun from Earth, Nietzsche means to alert the reader to the sheer gravity of the lost moral and psychological value systems which for so long were attributed to the structures of religious, specifically Christian, ideology in the 19th and 20th centuries.

However, the sharpness of this parable as critique is the dubious prospect of which direction humanity would now stray in the wake of God's death, if in any direction at all.

---

[1] Nietzsche, Friedrich. *The Nietzsche Reader*. Edited by Keith Ansell-Pearson and Duncan Large. Malden, MA: Wiley-Blackwell (2006), 224.

[2] Ibid.

[3] Ibid.

[4] Ibid.

This is found when the madman asks, “Whither is it [The Earth] moving now? Whither are we moving? Away from all suns? Are we not plunging continually?”[5] The ‘death of God’ was a metaphoric declaration that marked the annihilation of God as Western civilization’s overarching Truth-value *prima facie* and thus the moral structural-functions of religion along with it; worse still because it was of our own doing. Humanity killed God, according to Nietzsche. It is unclear, in the parable and elsewhere, what Nietzsche believed would replace God-as-Truth. Amid his lamentations over this sepulchral knowledge, the madman inquires to his audience about what newly cultivated value systems would usurp the religious morale. The madman goes on to ask:

> “How shall we comfort ourselves, the murderers of all murderers? What was holiest and mightiest of all that the world has yet owned has bled to death under our knives: who will wipe this blood off us? What water is there for us to clean ourselves? What festivals of atonement, what sacred games shall we have to invent?”[6]

It is difficult to assess the methods by which humans were responsible for annihilating God in this sense as Nietzsche failed to articulate this in his polemics. Apart from excoriating humanity (including himself) for this murder, Nietzsche insists that the greater task still awaits society. It is clear that by employing terms as ‘festivals of atonement’ or ‘sacred games’ he means to ask what new forms of moral—or, as my contention will be, ideological—values and their accompanying traditions will arise to replace God. He then suggests that perhaps this task of replacement would prove too great a task. In fact, when he asks, “Must we ourselves not become gods simply to appear worthy of it?”[7], the madman superimposes an ambiguous question on the part of humanity. He is cleverly suggesting, that in some form or another, humanity will be required to replace the God they killed by becoming an instantiation of him to prove

---

[5] Ibid.

[6] Ibid.

[7] Ibid.

themselves worthy of the deed. This ambiguity begs some questions. Was Nietzsche inviting humanity to create for themselves.

This death of God was essentially the subversion of traditional ideological monopoly in the form of staunch theological moral hierarchy with what can be accurately called intellectual secularization.[8] God did not die in the literal sense, our idea of one, nonetheless had perished. A general historical perspective is necessary to ascertain the significance of religion within political and social structures during the period when this declaration was made.

III. Ideology in the Secular

Beyond Good and Evil (henceforth *BGE*), perhaps Nietzsche's most well-articulated rejection of Western morality does well to establish correlates between the moral genealogy, history and the corresponding social behavior of the time. This line of inquiry was appropriate for addressing the nexus between geocentrism, anthropocentrism and the sociopolitical implications it had prior to the subversive scientific revelation of a heliocentric universe posited by Copernicus and Galileo. Earth was placed firmly at the center of the universe and so too were its inhabitants. Social hierarchies and the order in which society must follow was a determinism *fait accompli* of ostensible astronomical orderings. This anthropocentric dogma, once widely promulgated by the Catholic Church, crystallized the existential and social traditions preceding the enlightenment of the 18th-century.[9] According to the Church, humanity having been placed ostensibly at the center of the universe proved our dominion. Thus, divine anthropocentrism engendered meaning and purpose whilst justifying disproportionate social power. Although it was four years earlier that the declaration of the 'death of God' was made, it

[8] Matthew Mutter; *Culture and the Death of God*. Common Knowledge 1 September 2015; 21 (3): 512–513. Pg. 512

[9] Maria Pia Paganelli; We Are Not the Center of the Universe: The Role of Astronomy in the Moral Defense of Commerce in Adam Smith. *History of Political Economy* 1 September 2017; 49 (3): 451–468. Pg. 457

didn't seem that Nietzsche possessed robust solutions to the problem of religion and its relational behaviors until *BGE*. Nietzsche contends that faith and knowledge or, more specifically, instinct and reason acted in two discreet and still influential ways. One way in which these were manifested is as the inextricable elements, which far before the ascendance of Christianity, served as the *a priori* mechanism for moral valuation and, then, as a direct consequence, the framework for moral constraint.[10] Early on, Nietzsche finds discomfort in the modes of valuation which, according to him, remained an anthropological constant and found firm ground during the height of Greek rational thought; namely the Socratic equation: *reason = virtue = happiness*. That Greek rationalism and Christianity are themselves armed with differing axiomatic valuations is irrelevant for what Nietzsche took aim at were the very structures of moral valuation with which Greek rationality and, later on, Christianity would find their justification.

In the parable, a satisfactory answer is not given precisely as to how humanity had killed God. The origin of ideological belief is assessed in *BGE* as well when Nietzsche prompts any 'followers of history' to trace the evolution of scientific philosophy to that of the most pervasive processes of knowledge and understanding.[11] Nietzsche illustrates general ideological cultivation as a progression of hypotheses, fictions, valuations, and necessarily a will to believe. Emphasis ought to be placed on the aspect of ideological cultivation which surrounds the will to believe. Both Slavoj Žižek and Louis Althusser, in their respective accounts of ideology asserted the necessity of believe in ideological interpellation, that is, the molding of the subject by means of symbolic reification of fantasy and 'the real' in Zizek's account or systems of material-structural class domination in Althusser's. Althusser's structural account begins from a conception of an ideological edifice laden within society's social and political structures and institutions,

[10] Nietzsche, Friedrich *Beyond Good and Evil*. London: Penguin Books (1886), 113-114.

[11] Ibid. Pg. 115

what he aptly called Ideological State Apparatuses or ISAs. These apparatuses ranged from parochial, to political and associative to scholastic; through all of which, in their own independent way, dominant class ideology and the relations of production were crystallized and reproduced. Žižek's reading of ideology borrows largely from Lacan and Hegel and discusses ideology and subjectivity through the lens symbolic reality. It is important to note that Nietzsche would have likely opposed their accounts as both thinkers viewed ideology as irrevocable and mutually dependent on the subject. Nietzsche's tracing of ideological articulation, in large part having to do with true belief, is perhaps best summarized by Zizek's dissection of obedience when he says, "certainly we must search for rational reasons which can substantiate our belief, our obedience to the religious command, but the crucial religious experience is that these reasons reveal themselves to those who already believe"[12] and Althusser in equal measure when he asserts that the structure of ideology ensures " the absolute guarantee that everything is really so" that "if the subjection of the subjects to the Subject is well respected, everything will go well for the subjects: they will 'receive their reward'."[13] The capitalized Subject is the symbolic representation of the material account of ideology which Althusser denotes as the cite of ideologization, what Žižek, in his psychoanalytic account refers to as the 'Big Other'.

Religious traditions were, by nature, belonging to this genus of thought progression. After which, the social principles of empiricism and enlightenment thinking were afforded such privilege to the greatest extent. Empiricism, that is the contemporary understanding of the scientific method along with its concomitant methods for observation and systematization of knowledge in addition to enlightenment values which valorized secular humanistic values such as liberty, freedom, and free critical thought

---

[12] Žižek, Slavoj. *The Sublime Object of Ideology.* London: Verso (1989), 35.

[13] Althusser, Louis. *On the Reproduction of Capitalism*. London: Verso (1970), 197.

proved very quickly that political organization by divine right was illegitimate, and that widespread, consistent moral contemplation without reference to God was not merely possible but a more plausible endeavor than religious dependency and theological authority. Nietzsche did not explicitly anticipate this subversion in the parable; Nietzsche asked instead, "are we not straying as through an infinite nothing?"[14] This nothingness is a clear reference to Nietzsche's anxiety that after God's death, or the death of Truth in the Christian moral tradition, Western society would inevitably stoop into an abyss of nihilism whereby no exact Truth-value system could be given privilege; instead, all values would become devalued by virtue of relativist competition. What is important to note is that as contemptible as Nietzsche found Christian morality, he understood the corporeal benefits of an organized value system based on moral competency. Moreover, in *Twilight of the Idols,* Nietzsche forebodes the dangers of such exorbitant devaluation, "When one gives up the Christian faith, one pulls the right to Christian morality out from under one's feet…"[15]

IV. A Paradigm of Nihilism

The implications of God's death could be confronted in one of two ways. The absence of a dominant Truth-value system could be tackled head-on as one would do if one were a 'free spirit'. A 'free spirit', for Nietzsche, is an individual who feels awake at the dawn of God's death. This type of individual is one who actively imposes their own will to Truth and therefore takes up the gauntlet of their own judgement and power. One who rejects the standard moral valuations and instead cultivates for themselves subjective morals and reason. The 'free spirit' is often a recluse, one who frequently seeks a citadel far removed from the crowd. He seeks reprieve from the socially conferred Truth-values of the herd and recklessly subjects himself to the wild caprices of truth and morality, deforms it, and

[14] Ibid, Nietzsche 2006, 224.

[15] Nietzsche, Friedrich. *Twilight of the Idols*. Cambridge: Cambridge UP (2000), 58.

then conquers it.[16] The best indication of the construction of the *ubermensch* was in Nietzsche's towering work, *Thus Spoke Zarathustra.* Nietzsche reiterates on multiple occasions that the quality of individuality which constitutes the *ubermensch* was to any member of 'the herd', an act of punishment. To be a free thinker, exiled to uninhibited intellectual innovation and heresy was highly unfavorable.[17]. Far, fast, forgotten, and thrust loudly into a night without consequence that ends in the realization of more than you were ready for.

Those in the herd prefer much more to react to meaninglessness in the wake of God's death with a fashionable passivity. The cause Nietzsche described is known as passive nihilism. In other words, a nihilism characterized by the receding of the spirit; an implicit rejection of foundational societal moral or political principles without the subsequent productive or creative capacity to establish principles that were novel and substantive. I argue that this passivity complements the destructive psychical potential that Swiss psychoanalyst Carl Jung, described as *the shadow.* Jung contended that modern man was in danger of disregarding his own psychological potential for evil, as it were, his *shadow* and instead reflected it unto his neighbor. Jung saw the sociopolitical implications of this deference as potentially fostering animosities between alternate ideological postures when he says, "It has even become a political and social duty to apostrophize the capitalism of the one and the communism of the other as the devil, so as to fascinate the outward eye and prevent it from looking at the individual life within".[18]

It is worth elaborating on just how the Nietzschian concept of Passive nihilism and the Jungian concept of the *shadow* are complementary. Nietzsche probed deeply into the individual depotentiation and ideological proliferation under which lay the passive

---

[16] Ibid, Nietzsche 1886, 71.

[17] Ibid, Nietzsche 2006, 222

[18] Jung, Carl Gustav. *The Undiscovered Self*. New York: Signet Psychology (1958), 64.

nihilism that God's death had ushered in. Similarly, Jung believed that the individual's unconscious, irrational repression of the part of the ego, with all its potential for acrimony and malevolence, was still further darkened by a misguided commitment to the seemingly salvific qualities of the political ideologies of the day. I believe they worked in conjunction with one another to predicate the deification of ideologies and then the ideological antagonisms that precipitated the ideological wars fought during the 20th century. To my mind, the very depotentiation of the individual in her process of ideological possession was the unbridled response of society after the erosion of their traditional foci of moral valuation, that is, the Christian religion. Where Nietzsche's musings prove to be especially prescient is in the moments where the madman anticipates the 'infinite straying' and 'sacred games' that man's passive nihilism would've rise to. Suffice to say, these questions posed by the madman were the preliminary investigations into the symbolic and political realities that would possess humankind with the creation and adoption of these many social and political ideologies.

V. An Infinite Nothing

In the immediate wake of God's death, Western civilization indeed reacted in the latter form, with passive nihilism. The implications of God's inestimable death on contemporary society were the burgeoning of ideologies in the Western world during the 20th-century, whose misguided solution to killing God was merely to supplant him with secular conceptions *of him*. Nietzsche correctly predicted, that "given the way of men, there may still be caves for thousands of years in which his shadow will be shown."[19] The most influential ideologies to rise following Nietzsche's prophetic declarations and untimely death in 1900, were Bolshevism, social egalitarianism, fascism, and neoliberal capitalism. These are all social and political ideologies whose spectres preceded and whose sordid manifestations followed Nietzsche's works. Yet, the permeating influences

[19] Ibid Nietzsche 2006, 219.

they have had on society today occurred during the 20th-century. Many agree that these social and political ideologies constituted a new vanguard of values and beliefs. As Jonathan Reé reminds us, atheism is the new rule; one doubtlessly imbued with the militant certainty of empiricism, literacy, and materialist factuality.[20] But has God, as an embodied determination vanished? I say he has not. For instance, when describing the fundamental nature of the Bolshevik movement beginning in 1917, Bertrand Russell exclaims that it is not merely a political doctrine but also teems with similar flavors as that of a religion which possesses a set of elaborate dogmatisms and moral rigidities.[21] One finds it increasingly difficult to unmoor the genealogical identities of the ideological and theological *modus operandi* as being at once instinctively religious and psychologically obligatory. Indeed, for Žižek, the fundamental level of ideology was this very fantasy which valorizes our social reality. When he replaces the notion of the "illusion masking the real state of things" with the "(unconscious) fantasy structuring our social reality itself"[22] he is referring the ideological form of fantasy whereby individuals will "continue to walk as straight as we can in one direction" and where "we follow even the most dubious opinions once our mind has made up…."[23] Essentially, Žižek delineate the psychological process by which our beliefs/fantasies are bound to an unconscious practice of ideological deification, indeed where that deification is born in the fantasy itself.

Furthermore, in his book *The Road to Unfreedom*, historian Timothy Snyder employs the words of Vladimir Putin citing Russian political philosopher Ivan Ilyin, "A certain ideology dominated in the Soviet Union, and regardless of our feelings about it, it was based on some clear, in fact quasi-religious, values. The Moral Code of the builder of

---

[20] Rée, Jonathan *'Varieties of unbelief'*, Index on Censorship, 31:1, 2002, 192-198. Pg. 193

[21] Russell, Bertrand. *Bolshevism*: *Practice and Theory*. New York: Arno (1972), 9.

[22] Ibid Žižek, 30.

[23] Ibid, 92.

Communism, if you read it, is just a pathetic copy of the Bible."[24] In his analysis of the erosion of individual life, Jung believed, as Snyder's does, that the leaders of the mass state would inevitably become defied. He believed that the mass man would cling to the power of the state, all but "delivering himself up to it psychically as well as morally"[25] and asserting the reality that the "State, like the Church, demands enthusiasm, self-sacrifice and love."[26] When Lenin or Putin wished to make an injunction, they did so by close reference to their ideological progenitors. They deified themselves by evincing the ideologically deified. Bertrand Russell asserts that a true Communist is he who undertakes a set of ideological beliefs, which true or untrue, guide their utterances and actions.

Much of the social and political textures of the 20th-century were due entirely to ideological conflicts that are suggestive in the parable. If fascism, Nazism, and communism dominated large parts of Europe and Russia during this period, capitalism and egalitarianism as diametrically opposed embodiments of liberalism mobilized huge segments of the Western world then and does even more so now. It must not be forgotten that Truth-value systems act as the antecedent to psychosocial locomotion. The adoption of value systems is nearly always implicit and occurring at all times. With regard to the dominant social and political structures, as it were, Althusser's ideological apparatuses that were relatively independent and differentiated and within which ideology becomes reified, it is under these circumstances where individuals are already-always interpellated as subjects continuously practicing the "rituals of ideological recognition."[27] Žižek takes the concept of interpellation a step further. He argues that the fundamental essence of

---

[24] Snyder, Timothy. *The Road to Unfreedom: Russia, Europe, America*. New York: Crown Publishing Group (2018), 59.

[25] Ibid Jung, 26.

[26] Ibid.

[27] Ibid Althusser, 189.

ideology is the subject's oblivious relation to ideology as constituting the very essence of the social reality or rather "'*ideological' is a social reality whose very existence implies the non-knowledge of its participant to its essence.*"[28] As Arendt once declared, authority's "hallmark is unquestioning recognition by those who are asked to obey; neither coercion nor persuasion is needed."[29] The respective structural and psychoanalytic accounts, though opposed,[30] account for individual depotentiation with respect to interpellation, and even suggest its necessity.

That which we value motivate our thoughts; it takes place between the corridor of illusion and rational action wherein is a space that the tacit herd happily lay. We have grown so adulating of these ideologies. We desire the moral primacy and direction that they provide as we would a religious deity. Our inculcation in ideologies and the value systems they entail do not require us to accept their axioms as true, only that we accept them as necessary, at which point, as Žižek points out, they will reveal themselves to us as truth. Nietzsche condemned what this ideological embrace meant for humanity. He was acutely aware of the reluctance of the individual to derive, by his own intellectual capacities, genuine determinations of Truth and value after God's death.[31] Certainly, this awareness generated the skepticism seen in the parable when the madman question's the marketplace about the future of their moral landscape.

According to Jung, the idea of the Christian epoch was held to blame for modernism's areligious organization. The architecture of Christianity had schematized

---

[28] Ibid Žižek, 15-16

[29] Arendt, Hannah. *On Violence.* Boston: Mariner Books (1970), 45.

[30] It should be noted how significant the distinction between the structural accounts of ideology on one hand, and psychoanalytic Lacanian theory of ideology, espoused by Zizek are to one another. To discuss the manifold intricacies of these competing accounts is, however, beyond the scope of the paper. Despite this, the two theoretical frameworks often come to similar conclusions about the interpellation of the subject.

[31] Jenkins, Scott D. "Nietzsche's Questions Concerning the Will to Truth." *Journal of the History of Philosophy*, vol. 50, no. 2, 2012, pp. 265-289. Pg. 269.

our sociopolitical ideologies which themselves became defied. The Christian Logos had shifted to a secular one. Jung writes:

> Words like "society" and "State" are so concretized that they are almost personified. In the opinion of the man in the street, the "State," far more than any king in history, is the inexhaustible giver of all good; the "State" is invoked, made responsible, grumbled at….Society is elevated to the rank of a supreme ethical principle; indeed, it is credited with positively creative capacities.[32]

What Jung is describing here by using specific words such as "society" and "state" as it relates to the concretization of these structural concepts as ideological markers are Lacanian signifiers. This can be thought of as a visual or discursive sign which marks the relation of this sign's representations to a subject. According to Žižek, it is signifiers such as those used by Jung in his descriptions of personified language, which ties the subject to the signifier and initiates the process of subjectivation. The "crucial step", says Žižek, "in the analysis of an ideological edifice is thus to detect, behind the dazzling splendour of the element which holds it together ('God', 'Country' 'Party', 'Class'…), this self-referential, tautological, performative operation.[33] The madman is astute in his proximation. In addition to observing the murder of God as a concept, he also anticipated that due to the ways of men, the ideological edifice and its moral foundations and self-referential operations would merely be replaced with new ones.

It is the case that neoliberal capitalism and liberalism via consumerism and egalitarianism as modes of social production create and recreate themselves, aggrandizing themselves each time by their mere existence. By this I am referring to the reification of an index of liberal theories of political, social and economic rights as well as the ethos of equality emblematic of contemporary liberal democracy. Gilles Deleuze and Felix Guattari referred to this as desire-production.[34] This, to me, exists most

---

[32] Ibid Jung, 75.

[33] Ibid Žižek, 109.

[34] While I have used Deleuze and Guattari appropriately in this context, it should be known that there may be contention with this use as both thinkers were post-structuralists who were themselves skeptical about ideology as such.

perniciously in today's society where hyper-consumerism and the economization of all existing social spheres placed next to an ethos of equality and social-political rights have established a flow of desire, then bolstered its production by codifying it; an exponential and self-repeating process.[35] Capitalism and egalitarianism produce this desire but do so in two different ways. Egalitarian norms, for instance, can be imposed as a fundamental moral principle. It demands the position of an absolutist norm under which individual conduct and institutional arrangement ought to conform.[36]

Desire-production is more useful in delineating the specter of contemporary capitalist bloat. Capitalism grows and is maintained simply by the proliferation of production and consumption in the tangible sense and desire-production in the psychosocial sense where the more we consume, is the more we want to consume. Growth essentially catalyzes and instils in our machine minds more growth without awareness or imminent fear of plateau. The consequences of both ideologies are manifested in the externalities of environmental damage, populism, tribalism, and ideopolitical divisiveness, in addition to the presence of the *précarité* (the precarious worker or individual) under neoliberal capitalism.[37] On the one hand, the egalitarian notion instantiates a plane of moral value as an irrevocable moral authority and neoliberal capitalism on the other, like the production of unconscious, perennial repletion of material value without which we could hardly imagine or want to imagine our lives.

VI. Conclusion

The latter parts of the parable feature the madman reflecting on his premature declaration. He insists that the message his is promulgating, of the imminence of god's death and the

---

[35] Deleuze, Gilles and Guatarri, Felix. *Anti-Oedipus: Capitalism and Schizophrenia*. London: Penguin Books (1972), 140.

[36] Arneson, Richard, "Egalitarianism", The Stanford Encyclopedia of Philosophy (Summer 2013 Edition), Edward N. Zalta (ed.),

[37] Foster, Hal. *Bad New Days*. London: Verso (2015), 100.

implications it would have for the moral landscape, have fell on deaf ears. “My time is not yet. This tremendous event is still on its way, still wandering; it has not yet reached the ears of men.”[38] In one sense Nietzsche admonishes humanity for failing to heed his warnings for it implies at once an ignorance of our role in God’s death, and a resistance to rectify it by establishing substantial moral ballasts in his place. Hence, “the deed [creation of individuated value systems] is still more distant from them that the most distant stars—and yet they have done it to themselves.”[39]

As critics of individual depotentiation, it is likely that both Nietzsche and Jung would rail against structural explanations for interpellation as well as Lacanian psychoanalytic accounts *qua* Žizek. Indeed, both expositions of ideology are punctuated by a *necessity* of subjective interpellation. One that “always-already is”[40] and one that “in its basic dimension…is a fantasy-construction which serves as a support for our ‘reality’ itself: an ‘illusion’ which structures our effective, real social relation.”[41] Both men are essentially arguing that the prospect that one may escape from the throes of ideological possession is a null one. I doubt Nietzsche was fearful of either Althusser’s of Žižek’s interpretations of ideologization but rather understood more broadly how susceptible mankind have always been and still are to making Gods of all but themselves. An understanding, that it is clear, Jung had himself and expressed as much.

To my mind, this demonstrates not only the pervasive incognizance toward our passive nihilism but our unwillingness to embrace a more substantial and individualized method of Truth valuation and moral ascendancy. On this point, you find the convergence of Nietzsche and Jung both of whom derided what was effectively the depotentiation of the individual under deified ideology that would inevitably subsume it and society en

---

[38] Ibid Nietzsche 2006, 224

[39] Ibid.

[40] Ibid Althusser, 192.

[41] Ibid Žižek, 45.

masse. Alas, the madman has still come far too early. From the birth of rational virtue and theological absolutism to the mosaic of sociopolitical and socioeconomic ideologies that have shaken the Earth and command still our unremitting attention and subconscious participation, it seems that all we have learned from history is that we have not learned from history. I do not think the world was prepared for the death of God, nor do I believe they are in preparation even now. They are pitted too deeply into their false piety and pseudo-individualism and have done nothing to remedy the death of God but to recreate him innumerable times. Ultimately, it seems, the religious semblance is among the most naked of human cries. Yet, so long as humanity treats morality as a thing not to be possessed but as something that possesses us, we will continue to be ruled by our ideas; there will be no free spirits. As for me? My life is a real life, not some theological exercise, some enlightenment trip that has nothing to do with living.

Works Cited

Althusser, Louis. *On the Reproduction of Capitalism*. London, Verso, 1970.

Arendt, Hannah. *On Violence.* Boston: Mariner Books, 1970

Arneson, Richard, "Egalitarianism", The Stanford Encyclopedia of Philosophy (Summer 2013 Edition), Edward N. Zalta (ed.).

Deleuze, Gilles and Guatarri, Felix. *Anti-Oedipus: Capitalism and Schizophrenia*. London, Penguin Books, 1972.

Foster, Hal. *Bad New Days*. London, Verso, 2015

Jenkins, Scott D. "Nietzsche's Questions Concerning the Will to Truth." *Journal of the History of Philosophy*, vol. 50, no. 2, 2012, pp. 265-289.

Jung, Carl Gustav. *The Undiscovered Self*. New York: Signet Psychology, 1958.

Maria Pia Paganelli; We Are Not the Center of the Universe: The Role of Astronomy in the Moral Defense of Commerce in Adam Smith. *History of Political Economy* 1 September 2017; 49 (3): 451–468.

Matthew Mutter; *Culture and the Death of God*. Common Knowledge 1 September 2015; 21 (3): 512–513.

Nietzsche, Friedrich *Beyond Good and Evil*. London: Penguin Books, 1886.

Nietzsche, Friedrich. *The Nietzsche Reader*. Edited by Keith Ansell-Pearson and Duncan Large. Malden, MA: Wiley-Blackwell, 2006.

Nietzsche, Friedrich. *Twilight of the Idols*. Cambridge: Cambridge UP, 2000.

Rée, Jonathan *'Varieties of unbelief'*, Index on Censorship, 31:1, 2002, 192-198.

Russell, Bertrand. *Bolshevism*: *Practice and Theory*. New York, Arno, 1972

Snyder, Timothy. *The Road to Unfreedom: Russia, Europe, America*. New York: Crown Publishing Group, 2018.

Žižek, Slavoj. *The Sublime Object of Ideology. London*, Verso, 1989.

Russell Clarke is a 2022 graduate of Toronto Metropolitan University (formerly Ryerson) with a Bachelor of Arts degree in politics and governance with a minor in philosophy. His research interests include political theory with specific focuses on liberal theory, democratic theory, critical theory, ideology, and political and social ontology. He has published other works of political theory and philosophy in publications including Oxford Political Review, London School of Economics Undergraduate Political Review, and Areo Magazine. In addition, he has participated in summer fellowships and programs at Cornell and The Hudson Institute. He plans to pursue postgraduate degrees in political science in the hopes of becoming a lecturer and writer in political theory. In his spare time, he enjoys reading existentialist literature, getting tattoos, and going to the gym.

Moral Sanity Reformulated: Revising Susan Wolf's Sanity-Condition

Benjamin Edelson

I. Introduction

Some people think that moral responsibility is a metaphysical impossibility because the universe is causally determined. Others think that determinism must be false because we know *a priori* that free will (and therefore responsibility) exists. A third group sets itself apart from the first two in its rejection of determinism's relevance to the issue of moral responsibility at all. On this view, responsibility is made possible by certain psychological capacities, capacities which either exist or do not irrespective of the truth or falsity of determinism. So conceived, moral responsibility is *compatible* with a deterministic universe. The question of what exactly the pertinent capacities are, however, is the subject of ongoing debate among compatibilists. Several influential answers involve what Susan Wolf labels the 'deep-self view' – the idea one's will must be connected to some deep or ultimate manifestation of one's *self*. In her paper 'Sanity and the metaphysics of responsibility,' Wolf takes issue with the deep-self view, suggesting that there is a further condition to be met: *sanity*. Sane agents have the capacity to "cognitively and normatively recognize and appreciate the world for what it is."[42] Just as their empirical beliefs must reflect the world's physical reality, Wolf thinks, so must their values accurately reflect its moral reality.

Building on Wolf's critique of the deep-self compatibilists, I will offer what I think is a necessary revision to her so-called 'sane deep-self view.' While there is promise in looking to sanity as the necessary capacity for moral responsibility, I think Wolf errs in emphasizing the *substance* of one's moral values as a benchmark for sanity. This misplaced focus prevents Wolf's theory from being able to account for changes in genuinely thought-through values over time, as well as differences between values sincerely held by contemporary agents. As a result, it fails to accurately encapsulate our real-life responsibility-practices. A better articulation of the sane deep-self view would

---

[42] Wolf, Susan (1987). 'Sanity and the metaphysics of responsibility.' In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology.* Edited by Ferdinand Schoeman. Cambridge University Press, pg. 56.

focus more on the capacity to justify one's actions by citing *any* general behavioral principle, and less on the particulars of the principles themselves.

Wolf "embraces a conception of sanity that is explicitly normative."[43] This, I will argue, is her problem. My task is to conceive of moral sanity in a way that is *not* normative. In so doing, I hope to patch some of the sane deep-self view's holes and make it a more plausible candidate in the compatibilists' search for the capacity necessary for responsibility. I will first present Wolf's formulation of the sanity-condition. Then I will point out its problematic implications, suggest and defend my fix, and address potential issues with my proposal.

II. Wolf's Sanity-Condition

Wolf arrives at her conception of sanity via her dissatisfaction with Harry Frankfurt's view of responsibility. Frankfurt thinks that the distinguishing mark between agents and non-agents is the former's capacity not just to do as they want, but to critically reflect on those wants *and* structure their will accordingly. The capacity for 'second-order desires' – a desire "simply to have a certain desire"[44] – is not enough, for, as the author shows, there are agents who meet this criterion whom we would regard as poor candidates for responsibility. He offers the example of a 'willing' drug addict – someone who struggles against his addiction, but is not capable of caring "whether his craving or aversion gets the upper hand."[45] This addict, being "neutral with regard to the conflict between his desire to take the drug and his desire to refrain from taking it,"[46] lacks a capacity key to responsibility: the ability to "[want] a certain desire to be his will,"[47] or the freedom to

---

[43] Ibid, pg. 61.

[44] Frankfurt, Harry (1971). 'Freedom of the Will and the Concept of a Person.' In *The Journal of Philosophy*, Vol. 68, No. 1, pg. 10.

[45] Ibid, pg. 13.

[46] Ibid, pg. 12.

[47] Ibid, pg. 10.

"want what he wants to want."[48] Frankfurt labels this higher-level connection between one's desires and their will 'second-order volitions.' Responsible agents' "wills are within the control of their *selves* in some deeper sense"[49] – they are "not just psychological states *in* us, but expressions of characters that come *from* us, or that at any rate are acknowledged and affirmed *by* us."[50]

But the question remains: "Who, or what, is responsible for this deeper self?"[51] Why stop at second-order volitions? Why are not third-, fourth-, or fifth-order volitions necessary for responsibility? We are seemingly no more responsible for our second-order volitions than we are for our first-order ones.

Wolf answers by suggesting that, to really have second-order volitions, we must be able to direct our will in pursuit of the *correct* kinds of ends. For agents to "understand and evaluate their characters in a reasonable way, to notice what there is reason to hold on to, what there is reason to eliminate, and what, from a rational and reasonable standpoint, we may retain or get rid of as we please,"[52] they must possess "the ability cognitively and normatively to understand and appreciate the world for what it is."[53] *Cognitively* in that they can recognize a chair for a chair, and *normatively* in that they can recognize right from wrong. If agents are to correct their desires and wills in accordance with the world's normative makeup, their normative beliefs about the world must be correct – they must be *sanely* connected to the world. So, to be properly held responsible one's deep self need be sane.

---

[48] Ibid, pg. 15.

[49] Wolf, pg. 50.

[50] Ibid, pg. 49.

[51] Ibid, pg. 51.

[52] Ibid, pg. 59.

[53] Ibid, pg. 62.

Wolf thinks that this "explains why we give less than full responsibility to persons who, though acting badly, act in ways that are strongly encouraged by their societies...many male chauvinists of our fathers' generation, for example."[54] She acknowledges that it "would unduly distort ordinary linguistic practice to call...the male chauvinist even partially or locally insane," but, despite this, maintains that they indeed are insane in that the normative basis for their sexism is so terribly mistaken that it demonstrates a lack of capacity to grasp the objective moral makeup of the world. In that sense they are insane; they simply cannot appreciate reality.

Here a glaring question arises: "What justifies [Wolf's] confidence that, unlike the slaveholders, Nazis and male chauvinists…we are able to understand and appreciate the world for what it is?"[55] The debate between those who think ethical truths are objective and those who think they are subjective has a long history. But Wolf wisely avoids wading into that disagreement in any substantive way. Instead, she simply asserts that "nothing justifies this [confidence] except wide intersubjective agreement and the considerable success we have in getting around the world and satisfying our needs."[56] We will undoubtedly continue to revise and improve on our values going forwards. But it seems to her that we have a fundamental normative understanding of the world that Nazis and chauvinists lack.

III. Wolf's Problems

The first issue with the sanity-condition is its implication that, whenever wide intersubjective agreement about proper norms of behavior shifts (as it has over time, and no doubt will continue to), the conditions for sanity also shift. Wolf is confident that we are sane today, but by her criteria we could legitimately be called insane by the people of

[54] Ibid, pg. 57.

[55] Ibid, pg. 60.

[56] Ibid.

tomorrow. For example: according to some wide intersubjective moral agreement of the 1950s, contemporary chauvinists could be said to be genuinely self-correcting when they examined and reaffirmed their sexist values. Now, we find the chauvinists' introspection processes objectionable. We think they came to the wrong conclusion, and have our own thought-through reasons for believing this. In another 100 years, if wide intersubjective moral agreement shifts in favor of sexism, people might think the chauvinists were correct in their defense of their values. Such a shift, while perhaps improbable, is entirely plausible.

But the above means that in the 50s chauvinists *were* sane, are currently *insane*, and in the future they *will* be sane again. How could this be so if sanity is just the ability to recognize the world for what it objectively *is*? Surely the world's objective makeup has not changed since the 50s.

To ask this is not necessarily to argue against moral objectivity. It is merely to point out an inability to reconcile Wolf's standard of sanity – values endorsed by wide intersubjective agreement – with radical shifts in such agreement over time. For example: many people currently deeply disagree on the morality of euthanasia. Both camps have rigorous moral arguments for their respective positions. If, in 100 years, euthanasia is widely recognized as seriously unethical, then, on Wolf's account, the people of the future would be justified in regarding today's euthanasia-defenders as "unable [to] normatively recognize and appreciate the world for what it is" and therefore "not fully *sane*."[57]

But of course many euthanasia-defenders *are* sane. They are sane because they are capable of justifying their view by engaging in good-faith deliberation about how people should behave. Their sanity is not a function of which side of the euthanasia

[57] Ibid, pg. 57.

debate they fall on. Wolf makes the particular values one holds determinative of sanity, but is unable to provide any substantive test for which values are the 'sane' ones.

Consensuses also vary (radically) in different locations and cultures around the globe. Does *wide* refer to a given community, country, continent – or the entire world? Even if we could define the area, would we need 51%, 69%, or 82% agreement for a particular moral view to be 'objectively' sane? On the very contentious issues there is never 100% concurrence. And even on the less controversial ones there usually exist many different consensuses at a given time.

The second problem with Wolf's condition is that it eliminates the viability of genuine moral disagreement, which is a fundamental part of moral thought. On her view, whom may we validly hold responsible? Only, it seems, people who share our (objectively correct) values, but fail to live up to them. But this rather limited category does not include many types of agents we actually want to hold responsible. Wolf addresses this towards the end of her paper, admitting that her view implies "that anyone who acts wrongly or has false beliefs about the world is therefore insane and so not responsible for his or her actions."[58] For, "if sanity is the ability cognitively and normatively to understand and appreciate the world for what it is, then *any* wrong action or false belief will count as evidence of the absence of that ability."[59] She answers by suggesting that "typically, however, other explanations will be possible, too – for example, that the agent was too lazy to consider whether his or her action was acceptable, or too greedy to care."[60] Perhaps the agent has the capacity to recognize the objectively correct values, and so is sane, but simply fell short of acting upon those values because of other factors. In many cases, this response will suffice. But it will not help when we want

[58] Ibid, pg. 61.

[59] Ibid.

[60] Ibid.

to assign responsibility in cases where we *genuinely* morally disagree – in cases where both sides have indeed thought their values and positions through, and are committed to defending them. In fact, these cases are often the ones in which we are most desperate to morally blame.

The real trouble for Wolf's view arises in cases in which neither side is being sloppy, yet both are genuinely convinced that they are understanding and appreciating the world's normative makeup for what it is. "The suggestion that the most horrendous, stomach-turning crimes could only be committed by an insane person," Wolf writes, "must be regarded as a serious possibility, despite the practical problems that would accompany general acceptance of that conclusion."[61] The issue is precisely that in certain situations there is serious disagreement about what constitutes such crimes. To many anti-speciesists there is a 'Holocaust on Your Plate' every time you dig into a meal of steak (think 'MEAT IS MURDER!').[62] And yet there are other long-standing philosophical arguments explaining why eating non-human animals is morally permissible. Wolf's view implies that one camp is objectively morally insane. But anyone who has talked to thoughtful representatives from both these camps knows that is untrue. Ethical deliberation is difficult, and clearly-thinking people arrive at divergent conclusions. But this does not make them insane. If it did, we would have no way of knowing on which issues we currently hold sane or insane views – and yet Wolf insists that we are sane in most of our views.

To return to euthanasia: the opposing positions are marked by affirmations of two different moral judgments. Euthanasia-attackers endorse *A*: 'Life is intrinsically good, so one ought not kill.' And euthanasia-defenders endorse *B*: 'Life is good insofar as people enjoy it, so one ought not kill those who want to go on living.'

---

[61] Ibid.

[62] Hamilton, Jill (n.d.). 'Ethics Case Studies: Using the 'Holocaust' Metaphor.' *Society of Professional Journalists*.

Per Wolf, each camp should regard their respective opponents with a puzzling sort of moral indifference. 'We may genuinely disagree,' the attackers would be expected to say, 'but all that means is that in endorsing *B* you demonstrate an inability to grasp the objective normative makeup of the world. You are morally insane; therefore, it is unfair for me to hold you morally responsible for your actions when you enable people to commit euthanasia, even though they are knowingly committing '*horrendous, stomach-turning crimes*.''

No one would address their normative opponent like this. The euthanasia-attacker would actually say: 'We genuinely disagree, and your endorsement of *B* is mistaken for *x* reasons. You are morally wrong; therefore, I will hold you morally responsible for enabling people to end their lives.' For the attacker, the defender is a prime candidate for moral blame, precisely *because* they have the 'wrong' values.

That is why we want to hold Nazis, chauvinists, and slaveholders responsible. It is not because they hold the 'right' values, but fail to put them into practice – it is because they thinkingly endorse the 'wrong' values. This confusion is the reason for Wolf's distortion of "ordinary linguistic practice."[63] She correctly notes that philosophical reflection about words' meanings should be based in their "mundane,"[64] everyday usages, and claims her conception of moral sanity aligns with those conventions. Her argument, however, leads us to a picture of moral sanity that is undeniably contrary to those usages.

IV. Sanity Reformulated

For these reasons, Wolf's position needs some tweaking. We need a sanity-condition that does not lead to conceptually unacceptable conclusions, and more accurately describes our real-life assignments of responsibility.

---

[63] Wolf, pg. 57.

[64] Ibid, pg. 47.

Analogize ethics to a game. To hold your chess partner responsible for making good or bad moves, she must sufficiently understand the objective of the game, how the pieces move, etc. We would not hold someone incapable of grasping these rules responsible for doing good or bad things in the context of chess, because someone who lacked the capacity to understand the rules of chess would be 'chess-ly' insane. (This is a clunky term, but the point is made.) If your partner were unable to grasp the rules of chess and happened to make a poor move, she would not be deserving of chess-ly blame; if she happened to make a good move, she would not be deserving of chess-ly praise either. The feedback only functions if the receiver has a sound understanding of the system within which they are being blamed or praised. If the receiver does not understand the constitutive rules of the system, they are no longer operating within the system, and so we cannot evaluate them by the metric *of* the system.

So to evaluate people by a moral metric they must be capable of understanding and participating in the 'system' of morality. Wolf's sanity-condition allows us to evaluate by a moral metric only people who come to the 'right' moral conclusions, but of course we can use the metric to evaluate people who come to the 'wrong' conclusions as well – that's in large part the point of the metric itself. She thinks that if one is operating 'poorly' within the system, they cannot be judged by the standards of the system. But to be judged by the system's standards one just needs to be operating within the system in the first place. That is why moral sanity consists in the capability to understand the system itself.

The conditions under which valid moral feedback is given, then, will depend on what the 'game' of morality looks like. Offering a robust definition of morality here would exceed the scope of this paper, but the element that I think is key for my purposes is that morals exist in *codes* — codes of conduct.[65] They are principles that differentiate

[65] Gert, Bernard and Gert, Joshua (2002, rev. 2020). 'The Definition of Morality.' In *The Stanford Encyclopedia of Philosophy.*

between right and wrong behavior in a general sense, and are then applied to particular situations. Principles can be modified by other principles in certain complex situations, but they generally stand independently of any particular set of circumstances. Appeals to morally justify behavior, then, are appeals to abstract behavioral principles. Examples: 'act so as to bring about the greatest happiness for the greatest number;' 'pursue basic goods like life, knowledge, play, aesthetic pleasure, and sociability;' 'act in any given situation as the virtuous person would;' 'act only upon maxims that you can will to become universal laws.' Moral decisions are made by applying general rules like these to individual situations – they are never made arbitrarily, for they must be justifiable if questioned.

For beings to be candidates for moral feedback they must be capable of understanding this. They must be capable of recognizing the project of ethics for what it is: the task of formulating correct *principles* of action. Whereas only certain people play chess, and only for a given amount of time, everyone is always 'playing' the game of morality, for we are all constantly behaving.

The root of Wolf's difficulties is her offering too *narrow* a conception of sanity. Compare the euthanasia-attacker and defender, who both have a proper understanding of moral thought, and engage in good-faith attempts to justify *A* and *B*, with a young child. The child comes to a conclusion about what should be done simply on the basis of her emotional, one-time response to the situation – perhaps death upsets her greatly, so she says that the physician shouldn't help end the patient's life – and is therefore unequipped to grasp the nuance of moral thought. She cannot grasp the abstract prescriptive force of the moral arguments at play – perhaps she cannot understand what is meant by 'intrinsic' vs. 'instrumental' goods – and so can only justify her behavior on a moment-to-moment basis.

This tension between what one may *want* in the present moment and what they think is *right* in general is a hallmark of moral thought. No doubt George Washington didn't want to get in trouble for chopping down the cherry tree, but this impulse was overruled by the power of the general prescription that one should not lie. One's momentary desire may often align with one's values – but the capacity to recognize that, and act on the desire *because* it aligns with one's principles, and not merely because one desires it, is what makes for moral sanity.

I would therefore reformulate the sanity-condition as: *the capacity to justify one's actions by appeal to general principles of behavior*. And since values are merely general behavioral principles that prescribe the pursuit of something of value, we may say in even simpler terms that to be morally sane one must be capable of justifying her actions by appeal to values. When one is capable of thinking through which abstract *ought*-principles she subscribes to, she is morally sane, and thereby an appropriate target of moral praise and blame. So conceived, moral sanity is broad enough to leave room for both shifts in values over time and genuine moral disagreement between contemporaries. We may disagree with someone, but if her justifications have moral integrity, we tend not to label her insane. Only if she is incapable of formulating her values as principles – incapable of formulating an ethical argument – is she morally insane. This description of moral sanity both is internally consistent, and more fully captures the way we actually assign responsibility.

V. Defending the Reformulation

It might be said that the picture of morality I have proposed is too broad, and that morality is just about doing the *right* thing, not any thing that one might be able to justify by appealing to general behavioral principles. But to think like this is to fall into the trap that defeated Wolf. When I refer to someone who 'thinks morally' I refer to someone who is *capable* of moral thought, not necessarily someone who arrives at the 'right' moral

conclusion. Someone can still think through what they should do in a given situation and come to a poor conclusion via poor values and/or empirical considerations. But they are still capable of moral thought, albeit poor thought – and, per compatibilism, it is the relevant capability that I am trying to accurately describe. All I have said is that if one can justify some behavior *P* by explaining why the reasoning underlying *P* holds in other situations as well, and not just in the current situation, then they are capable of moral thought, and are therefore proper targets of moral feedback. This is not overly broad.

At times principles of action conflict, and we are hard-pressed to decide between them. We seem to have an evolutionary 'soft spot' for entertaining values that empathetically consider others' interests, since teamwork greatly aids survival prospects. Perhaps, however, there do exist some cases in which it is more correct to disregard these interests wholly in favor of one's own. The ethical egoist thinks so. And there are plenty of other points of disagreement: there are virtue ethicists, hedonistic utilitarians, preference utilitarians, deontologists, natural lawyers, new natural lawyers, feminist ethicists – the list goes on. The disagreements between these camps concern the particulars of moral theorizing and action – but regardless of the particulars of their plans of action, they all justify their plans by appealing to general principles.

It might be objected that my reformulation overly focuses on agents' capacity for principled, rational action, and fails to mention some capacity for emotional sensitivity to the suffering of others. In his paper 'The Conscience of Huckleberry Finn,' Jonathan Bennett shows how "sympathy" can act as an important counterbalancing force in people who arrive at a "bad morality"[66] purely deliberatively. In freeing Jim, Huck acts in accordance with his passions – his emotions – and against his principles. If we think Huck is a valid target of moral praise, isn't emotional sensitivity sometimes a necessary condition for responsibility?

---

[66] Bennett, Jonathan (1971). 'The Conscience of Huckleberry Finn.' In *Philosophy*, Vol. 49, pg. 1.

Often emotional sensitivity will make for a 'good' agent. But we are interested in the conditions necessary for responsible agency itself. And there are some agents who lack real empathy that we would hold responsible. Consider the ethical egoist, who thinks that she "morally ought to perform some action if and only if, and because, performing that action maximizes [her] self-interest."[67] Perhaps the egoist feels sympathy for others; perhaps she doesn't. Regardless, we will want to hold her responsible when it becomes clear that she has the capacity to justify and act upon volitions we find objectionable. The reasons for this are identical to the ones presented in the suicide example. Making emotional sensitivity a necessary condition for responsibility would lead us to the same problems that Wolf's sanity-condition did. We often hold emotionally insensitive people responsible insofar as they've thought through their values. The sanity-condition must be broad enough to hold responsible people acting in accordance with a variety of behavioral norms, and numerous norms eschew emotional sensitivity. Emotions constitute a unique aspect of moral thought, and play an important role in moral psychology – but I don't see them as necessary for moral *responsibility*.

There are different reasons for not exercising the relevant capacity as I have described it: some people simply don't have it, others have it but it is underdeveloped, and others still have it yet willfully do not engage it. The second category could refer to someone who has unquestioningly swallowed the values of their society and never arrived at their own normative formulations. It could also refer to an adolescent who is in the process of developing the capacity. Our actual praise- and blame-bestowing practices confirm that we treat these two cases somewhat similarly – they are cautiously deserving of *some* responsibility, but not in the robust way that a fully morally rational adult is.

Huck seems to fall into this category; his capacity for moral thought exists in some basic form, but is critically underdeveloped. In rejecting his principles he begins to

---

[67] Shaver, Robert (2002, rev. 2021). 'Egoism.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.

thoughtfully reflect on what to do: he mulls over the circumstances in detail and agonizes over the conflict between his "general moral principles and particular unreasoned emotional pulls."[68] But he ultimately decides that, since he will feel bad either way, going forwards he will "do whatever 'comes handiest at the time' – always acting according to the mood of the moment."[69] This is the mindset of a child, a being that is driven by whims and shys away from confronting uncomfortable difficulties through open deliberation, of someone who refuses to search for principled justifications for their behavior and so is inconsistent in the quality of their actions – someone incapable of real moral thought. Huck's mistake is his failure to revise his principles on the basis of his sympathies; if he had done that, he would be a fully responsible agent. But he lacks the ability to engage in the "abstract intellectual operations"[70] necessary to effectuate that revision, and so decides to do away with principles altogether. He is therefore a less-than-clear case; perhaps he is deserving of *some* responsibility.

As for those who have the capacity and willfully do not engage it: if your chess partner who has the capacity to understand the game makes a stupid move, you would probably still hold her chess-ly responsible. This is because, if asked, upon reflection she could provide a satisfactory explanation of why her move was poor and what a better move would have been. So in holding her responsible you would be, in a certain sense, accusing her of not living up to her potential. Some chauvinists of the 50s, to return to Wolf's example, are therefore appropriate candidates for blame, depending on their capacity to justify their sexist beliefs. Others are not. Another situation in this category might be someone who performs an immoral action under pain of death. Many philosophers of responsibility have tried to show that such a person is not responsible

---

[68] Bennett, pg. 4.

[69] Ibid, pg. 8.

[70] Ibid.

because her will is not free, or she could not have done otherwise, or some other reason of the like. I think this way of approaching this situation is mistaken. *If* the person in question can justify her self-preservatory actions by citing some formulation of the principle 'one ought to, or is at least justified in, valuing the perpetuation of her own life above more trivial moral ends,' she is a responsible agent. She is not not responsible for doing as she did – she thought through her action, and performed it – but her adherence to the principle of self-perpetuation makes it inappropriate to fully blame her for her action. In fact, someone like the egoist might even think her deserving of moral praise. On the other hand, if the person in question cannot justify their actions by appeal to such a principle, then she is not responsible.

How are we to know if someone has the relevant capacity and is not exercising it, or simply doesn't have the capacity at all? This is an important question for all compatibilist theories, not just mine. I answer: ask the agent. If they can provide a general principle explaining why they *ought* to have behaved in the way that they did, then they are morally sane, and so an appropriate candidate for moral praise and blame. If their reasoning is incoherent, or they cannot provide any morals by which to justify their behavior, then they are not an appropriate candidate because they are morally insane.

Psychopaths are an interesting case. It is unclear if they act according to a generalized schema about what is good for themselves, like the egoist, or if they are really just impulsive (i.e., lack sane second-order volitions). The former would be a valid target of praise and blame; the latter would not. No doubt there is some variation – and, resultantly, inconsistency in the definition of psychopathy. Psychopaths do not really undercut the intuitive appeal of my conception of moral sanity/responsibility, I don't think, because our responsibility-practices are complex. There is disagreement about how to handle some agents. Compatibilism's ability to account for hazy cases like these is part of its appeal. The idea that the moral capacity is something that must be developed is an

old one; Aristotle thought that moral character developed only over time and by a familiarity with practical ethical situations.[71] We become responsible agents as we come to a full understanding of what moral thought *is* through regular exposure to situations in which people praise and blame us, and as we become capable of critically reflecting on that praise and blame's appropriateness (its accordance with general principles of action). This development takes place most crucially throughout childhood and adolescence;[72] no doubt it continues through adulthood as well. The question of whether psychopaths experience this development seems an empirical one, and not one I am prepared to take on here.

My formulation of moral sanity is purposely broad enough to encompass *all* moral judgments. It is important to stress that in this broadening I am not endorsing moral subjectivism, nor arguing against objectivism. I am not implying that there cannot be correct or incorrect judgments; the realm in which this paper is operating is one step removed from evaluating any particular moral judgment. It is concerned with figuring out what counts as a valid moral judgment in the first place, and arguing why the capacity to properly justify these judgments is what constitutes moral sanity. The fact that one arrives at a particular judgment, correct or incorrect, is not sufficient grounds for labeling them insane. There are more relevant pieces of the puzzle.

VI. Conclusion

As reflective creatures – creatures capable of second-order volitions – it behooves us to come to our own moral conclusions. These conclusions will often be contrary to wide intersubjective consensus, but that is not a bad thing, for exposing our beliefs to criticism (both the criticism of popular opinion, and our own) only strengthens them. To do this

---

[71] Homiak, Marcia (2003, rev. 2019). 'Moral Character.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.

[72] Fine, Cordelia and Kennett, Jeanette (2004). 'Mental impairment, moral understanding and criminal responsibility: Psychopathy and the purposes of punishment.' In *International Journal of Law and Psychiatry*, Vol. 27, pg. 425–443.

properly, Wolf correctly notes, we must be morally sane. Her sane deep-self view satisfactorily answers the problems that defeat Frankfurt's 'plain' deep-self view. But her formulation of sanity leads to conceptually unacceptable conclusions, and in key cases doesn't match up with our real-life responsibility-practices. In its explicit normativity, her view fails to leave adequate room for genuine moral reflection. Reformulating sanity as the capacity to engage in this reflection, I think, strengthens the sane deep-self view greatly.

Works Cited

Wolf, Susan (1987). 'Sanity and the metaphysics of responsibility.' In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Edited by Ferdinand Schoeman. Cambridge University Press.

Frankfurt, Harry (1971). 'Freedom of the Will and the Concept of a Person.' In *The Journal of Philosophy*, Vol. 68, No. 1, pg. 5–20.

Hamilton, Jill (n.d.). 'Ethics Case Studies: Using the 'Holocaust' Metaphor.' *Society of Professional Journalists*.

Gert, Bernard and Gert, Joshua (2002, rev. 2020). 'The Definition of Morality.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.

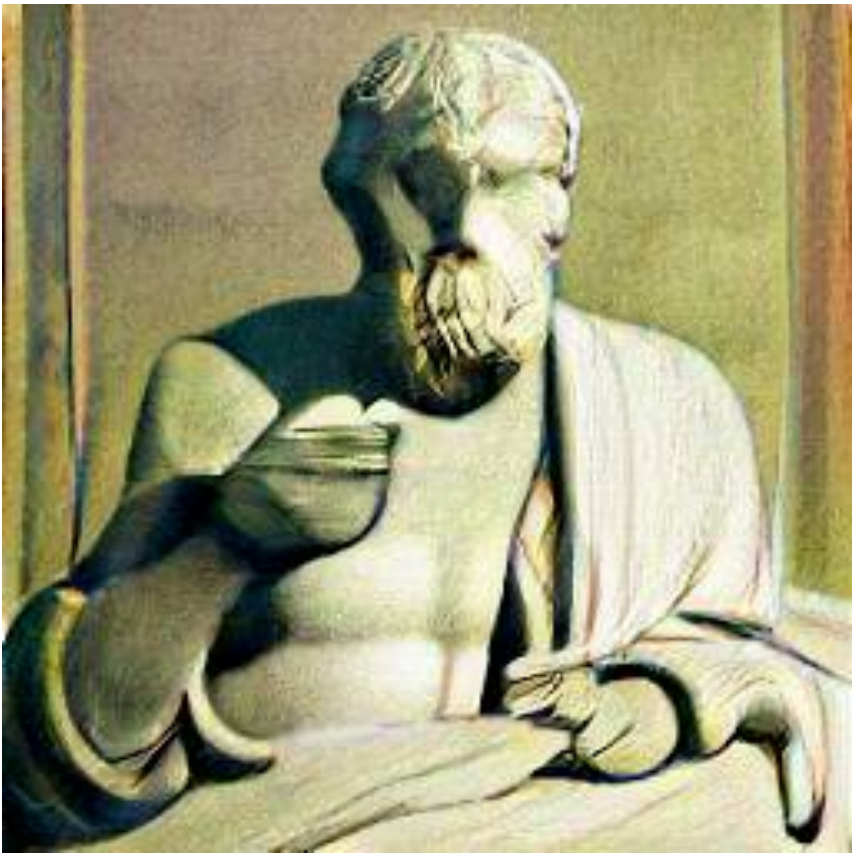
Homiak, Marcia (2003, rev. 2019). 'Moral Character.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.

Fine, Cordelia and Kennett, Jeanette (2004). 'Mental impairment, moral understanding and criminal responsibility: Psychopathy and the purposes of punishment.' In *International Journal of Law and Psychiatry*, Vol. 27, pg. 425–443.

Bennett, Jonathan (1971). 'The Conscience of Huckleberry Finn.' In *Philosophy*, Vol. 49, pg. 123–134.

Shaver, Robert (2002, rev. 2021). 'Egoism.' In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta.

Kantian Practical Ethics is Empty



Anson Berns

Abstract

I argue that Kant's deontological ethical theory espoused starting with the *Groundwork for the Metaphysics of Morals* is empty of practical ethical content. I detail approaches towards formalizing its practical content through a decision process for the categorical imperative (a CI-Procedure) and the problems with such an approach. Also considered are attempts by Kantians to endorse a version of the theory with minimal or no practical content, as well as how Kant and Kantians present themselves as applying their theory to practical questions. I discuss implications of this emptiness, and argue that it is a serious problem for the Kantian project as it sees itself.

The meat of Kant's *Groundwork for the Metaphysics of Morals* is focused on topics that would now be considered part of the domain of metaethics: sources of obligation, how morality binds us, what it means for an action to be "right," etc. When Kant discusses how to apply these principles to practical ethical questions, however, he is much less clear. It is so obscure, in fact, how to translate Kantianism into a practical ethics that I will argue here that in practice there is in fact no practical component to the theory (or in other words a multitude of different kinds of first order ethical reasoning are all compatible with Kantianism.) In this way, the usual assumption that Kantian deontology is a direct competitor to, for example, utilitarian consequentialism is revealed to be an illusion.

The contention that Kant's ethical theory is devoid of practical prescriptions is not a new one. Franz Brentano wrote in the 19th century that a serious problem with the categorical imperative was that "even if one were to accept it, one could not use it to deduce any ethical consequences" (Brentano 31.) Even earlier, Mill criticized Kant on the grounds that "when he begins to deduce from this precept [the categorical imperative] any of the actual duties of morality, he fails, almost grotesquely…" (Mill 9.) Most of these criticisms, however, are not particularly fleshed out and take as their targets the obscurity and ad hoc quality in which Kant discusses practical questions. Instead of merely criticizing Kant for how he applies the principles he lays out to practical situations, I will endeavor to provide a positive argument for why Kantianism cannot have the practical content we want from it, and explain the consequences of this fact.

To do this, I will first outline the classical attempts to formalize the application of the categorical imperative to real life situations (the "CI-Procedure.) I will then discuss the problems with such an approach, both how it is not successful on its own terms (because of difficulties with "puzzle maxims") and how it strays from Kant's original spirit. With these considerations in mind, I will consider how we might still obtain

practical guidance from Kant's theory and analyze how Kant and Kantians take themselves as doing so. Finally, I will conclude that the genuine practical content of the Kantian theory is minimal, and that this is a problem because it is incompatible with the untenable way in which Kant and Kantians claim to apply their theory practically.

In what follows, my focus is directed on Kant's first formulation of the categorical imperative, particularly its variant known as the Formula of the Law of Nature: "Act as if the maxim of your action were to become by your will a universal law of nature" (*G* 4:421.) This is both because the literature on this topic (particularly on puzzle maxims and CI-Procedures) focuses on this formula, and because the broader points about Kantian practical ethics can for the most part be made equally well regardless of formula, but are made clearer if a specific one is chosen for the categorical imperative.

To start, let us look to the most straightforward and classic attempt to formalize how Kantian practical ethics would work. What we would like is to be able, when confronted with a hard problem, to discern what the categorical imperative binds us to do in this particular situation. We would like to be able to do so without having to be particularly inventive in our arguments or read Kant's mind. In other words, what we want is a decision procedure that, given a situation, outputs our relevant obligations. This is referred to in the literature as a "CI-Procedure" (CI being categorical imperative) and Rawls's classic presentation breaks it down into four steps. First, one formulates one's maxim of action as "I am to do X [action] in circumstances C in order to bring about Y [state of affairs]" (Rawls 83.) Then, the second and third steps transform this maxim into the universal law of nature "Everyone always does X in circumstances C in order to bring about Y." Finally, one analyzes the world in which that law is added to the extant laws of nature and determines whether it is possible to will it (or even to conceive of it.) This

procedure does in fact mirror the way Kant applies the categorical imperative, at least sometimes (see for example *G* 4:421-3.)

The most immediate difficulties in actually implementing this procedure lie in the determination of what C and Y are for a particular situation, and the final step of analysis of the hypothetical world. These difficulties can be brought to bear when considering so-called "puzzle maxims" which give unintuitive results when passed through the procedure. Consider first this example from Timmermann: "A maxim along the lines of 'I want to dine at a friend's place at 7.00 pm on Mondays' cannot be universalised if we assume that the particular friend in question must be present, for example to discharge his or her responsibilities as the host of the party [since he too would be dining at a friend's place]" (Timmermann 157.) Thus, the CI-Procedure rejects this action even though we would typically think it obvious that there is nothing wrong with it. The obvious resolution to this problem is the claim that this maxim is an inappropriate choice for the action. After all, many maxims could describe the action and the level of detail to our choice seems odd. Certainly, other maxims describing the same action (such as "I want to enjoy the company of friends") get through the CI-Procedure just fine, and so the worry here should not be that the categorical imperative stands against us dining with our friends at 7:00 pm on Mondays.

Rather, the challenge lies in explaining exactly why the aforementioned is the wrong maxim to choose. Some solutions see the problem in the level of specificity of the maxim itself. Bittner suggests that maxims are supposed to be "rules of life" and thus should have more generality. However, this is both vague and unrepresentative of Kant's own usage. Timmermann gives the example of the suicidal man's maxim to end his life "when its longer duration is likely to bring more pain than satisfaction" (*G* 4:422) as a maxim tested by Kant against the categorical imperative that is relatively specific and would serve as a bad "rule of life." Timmermann's preferred solution is to instead

emphasize the connection between maxims and ends. Choosing the appropriate maxim for a given action should be guided by the ends we have in mind when taking that action (represented as Y in Rawls's version of the CI-Procedure) since "a maxim is more than just an action-guiding rule; and doing something as a matter of principle, because one is directly interested in it as an end, is relevantly different, morally, from doing it merely as a means towards some other end" (Timmermann 158.) This approach seems promising, but it involves conceding that if someone, for some reason, wanted to eat at 7:00 on Monday with her friend for its own sake, that would in fact be morally wrong. Timmermann does concede this, "someone who has developed a bizarre inclination to dine with his friends at 7.00 pm on Monday nights *as such and under that description*… ought not to do so, precisely because his maxim would fail the test of the categorical imperative" (Timmermann 158.) Conceding this point seems either circular or preposterous. Although such a person would definitely be "bizarre," it seems equally bizarre to say that someone with such a fixation is morally barred from acting on it because it fails the CI-Procedure.

Another classic puzzle maxim comes from Brentano. He asks us to imagine a civil servant who is offered a bribe and takes it, because of the categorical imperative. In his words, "if the contrary maxim [I will not accept bribes when asked] were to become a universal law, then people would no longer attempt bribery" (Brentano 31) and thus that hypothetical world could not be willed. This reasoning is strikingly similar to Kant's own reasoning about why keeping one's promises is a duty. A maxim of promise-breaking, if universally willed, would destroy the institution of promises altogether creating a contradiction, in much the same way that a maxim of bribery-denying would destroy the institution of bribery altogether creating a contradiction. Since bribery denial seems to be obviously not morally wrong (in fact, it seems morally required) the CI-Procedure has seemingly failed here. One might argue that the way in which the hypothetical world was

analyzed was flawed and that the undermining of bribery can be willed for (perhaps despite the contradiction) unlike the undermining of promises. Or, perhaps, one might again insist that the maxim was wrongly chosen here. That is Timmermann's approach once again, who declares, again stressing the connection between maxims and ends, that "turning down bribes is a means to a legitimate end, but it should not be considered worth doing for its own sake" which means "the civil servant's practical principle must be a general maxim of decency" (Timmermann 159.) This again feels very ad hoc. When can maxims be individually tailored to the situation at all and when must they be general maxims of decency or indecency? Would it have been possible to declare in this case that the maxim must be general without prior knowledge that the specific maxim would wrongly fail the universalization test?

A final puzzle maxim worth considering is one in which someone (perhaps a Nazi or similar right-wing ideologue) has as a maxim the abhorrent intention "I will eliminate all members of inferior races." This can easily be universalized, though it obviously cannot be encouraged or permitted by the categorical imperative. The problem in this situation lies not in determining a level of detail or generality in the maxim, but rather the fact that the maxim incorporates the proposition that certain races are inferior. It feels as though this non-factual content of the maxim (non-factual both in that it is false but also in that it is of an evaluative nature) should not be allowed to be subject to the generalization test. One might argue again here that the problem lies not in the initial step of determination of the maxim, but rather in the ultimate analysis of the hypothetical world with it added as a law of nature. Perhaps there is some argument that we cannot will such a world after all, despite it containing neither obvious contradictions or obvious states contrary to the agent's self-interest, the foundation of Kant's previous arguments that worlds were impossible to will.

None of these cases alone is insurmountable for a proponent of the CI-Procedure, but each one adds in a constraint that she must account for. One could, as Timmermann does, that puzzle maxims do not represent the correct maxims in the situation, or like Bittner that they are not really maxims at all. Or one could redescribe the calculation of whether the ultimate hypothetical world is willable. In any case, it is clear that the CI-Procedure has large holes that, if filled in an ungraceful and ad hoc manner, cast significant doubt on the theory.

Perhaps the CI-Procedure can be further specified to account for all sorts of puzzle maxims and edge cases, and perhaps it cannot. Either way, the additional baggage added onto the principle seemingly produces doubt of this approach on its own. It also becomes quickly unclear what the exact source in Kant's body of work is supposed to be —or if the clarifications are supposed to be unsaid logical consequences of the work their sources are obscure as well. Without delving deeply into how Kant justifies both the categorical imperative in general and our chosen Formula of the Law of Nature for it in particular, there is an obvious incongruity between the nature of Kant's arguments and any total description of such a 'complete' CI-Procedure. Kant's arguments about the metaphysics of morals rarely enter a mode of description either sufficiently detailed or sufficiently clear to define such a low-level practical procedure. He is more concerned with sources of moral motivation and the grounding of how it binds us.

This critique on grounds of drifting away from Kant's purer metaethical character is much in line with what Kantian Allen Wood says on the matter. Wood is less concerned with justificatory gaps in the detail of the CI-Procedure *per se* but instead its discontinuity with the original spirit. Wood argues that such an interpretation misunderstands what a moral principle such as the categorical imperative is for. He contends that constructing a CI-Procedure at all takes for granted "that moral philosophy is concerned solely with solving intellectual problems about the rational procedures to be

used in making decisions and justifying them" (Wood 15.) Wood thinks that rather than needing to be equipped with a decision procedure, a moral agent merely must have the "intellectual capacity to distinguish right from wrong" along with "the strength of character and the good judgment to do so" (Wood 18) and that this framework should supplant any CI-Procedure based framework. Note that under this interpretation, puzzle maxims immediately become significantly less puzzling. Wood says to those who see the situation like Timmermann that "those who reply to these counterexamples by saying: 'this isn't the agent's real maxim' [wrongly keep] the persisting pretense that FUL/FLN can after all be used as general tests for the permissibility of maxims after the manner of a 'CI-Procedure'" (Wood 33.)

For Wood, the categorical imperative provides "moral orientation" and perhaps keeps one generally on the right track, but right action is a fundamentally judgmental, rather than intellectual, activity. Wood sees the value of this "moral orientation" (Wood 18) as being a reminder to never make exceptions from duty for oneself. Wood approvingly quotes Kant's insistence that it "with this compass [the categorical imperative] in hand, [common human reason] knows its way around very well in all the cases that come before it" (*G* 4:403.) This interpretation returns duty to the central role in Kant's theory, relegating maxims to useful theoretical constructs. The role reason plays in this theory is confined to that of common sense. Seemingly, this interpretation both jells with the spirit of Kant's work and eliminates the serious problems that plague the CI-Procedure.

Surely, however, we must be able to recover *some* first order component to the theory, though. Just because we emphasize duty and de-emphasize reasoning in the Kantian program does not mean that moral prescriptions for specific situations are *never* a direct consequence of the categorical imperative. If we go too far overboard in a project of refocusing Kantianism on the metaethical, we risk absurd conclusions, like that

Kantian ethics and utilitarianism are compatible. If they are then it seems as though the practical constraints imposed by Kantianism are so minimal as to be useless in determining morals at all. So far, these arguments are merely sketches since perhaps Wood or someone like minded might be tempted to deny the problem a lack of practical content poses for Kant. I will consider more later the challenges Kantians biting the bullet on this question face, in particular when faced with how Kant and followers actually discuss moral situations. First, however, we must establish how some first-order content might still survive our Wood-style elimination of CI-Procedures.

One question, given the above consideration, is how we can logically eliminate the undesirable thesis that Kantian and utilitarianism (or consequentialism more generally) might be able to logically fit together. Given the leeway already established, we might even think of how such a synthesis might look: a consequentialist (of whatever flavor, those details could be filled in) whose moral code is justified thusly: "My maxim of action is always to do what produces the best outcome for everyone. If I ever acted elsewise, my maxim would be one which prefers worse outcomes to better ones, and therefore I could not will it to be a universal law since everyone acting that way would be bad for the people as a whole, a group of which I am a part." Presumably this cannot be a valid application of the categorical imperative—or if it can then a synthesis of Kantianism with almost any coherent moral theory can, stripping away any real normative content from the categorical imperative once and for all. But what *precisely* has gone wrong for the Kantian-consequentialist, if anything?

First, one might take issue with the characterization of the maxim of all actions that are not consequentially optimal being a preference for worse outcomes, but in fact this seems like a fair assessment when we reframe the problem around duty. If the duty in question is the duty to promote the outcome that is best for everyone (i.e. that is optimal in maximizing good consequences), then it is reasonable to characterize any non-optimal

choice as making an exception for oneself from this duty and being driven by a maxim that one should act non-optimally, perhaps under certain particular circumstances.

One might also take issue with the form of the reasoning about why this is not universally willable. My welfare being non-optimal is not guaranteed from people acting in such a way as to not promote the optimal general welfare. However, it is quite likely. The cases in which we might wish that those around us did have non-consequentialist maxims are only those in which our welfare would be sacrificed for a greater increase in the welfare of others. At first, this exception seems incredibly significant, but consider Kant's reasoning for why the categorical imperative impels us to be charitable and sympathetic to others: "a will that decided [that a maxim of non-sympathy should be a universal law of nature] would conflict with itself, since many cases could occur in which one would need the love and sympathy of others and in which, by such a law of nature arisen from his own will, he would rob himself of all hope of the assistance he wishes for himself" (*G* 4:423.) Here, Kant neglects the fact that the burden of having to be charitable to others may outweigh the lack of charity given to oneself, especially if one is in a privileged position. The contradiction is merely that in "many cases" one would be forced to will that they not be given assistance (and at the same time through common desire will that the assistance be given.) It seems that a similar thing can be said of the consequentialist version, then. Just as in the case of sympathy, "many cases could occur in which one would need" others to act in a way to optimize the consequences, since that optimization would include optimization of consequences for myself.

In fact, the pattern of reasoning that our hypothetical Kantian-consequentialist uses seems to mirror very closely Kant's application of the categorical imperative to the duty of beneficence to others. Indeed, the duty to "contribute anything to his welfare or to his assistance in need" (*G* 4:423) seems to be itself of a consequentialist form. Of course, this does not collapse Kantianism into consequentialism, but it does eliminate an

objection to the above consequentialist application of the categorical imperative that doubts whether duties can have such a consequentialist character. If there is a duty to contribute to the welfare of others, from what then can Kant derive his opposition to consequentialism? Plainly, the answer lies in the resolution of conflicting duties. If there is a duty with a consequentialist form (the duty of beneficence), but I am not impelled by duty considered as a whole to be a consequentialist, then it must be that in many cases some other duty is what binds my action.

Kant was very unclear about what to do in situations of conflict. In the *Metaphysics of Morals*, he denies the possibility of conflicts between duties or obligations, saying "since duty and obligation are concepts that express the objective practical necessity of certain actions and two rules opposed to each other cannot be necessary at the same time… a collision of duties and obligations is inconceivable" (*MM* 6:224.) However, he does admit conflicts between the *grounds* of obligation, saying that "when two such grounds conflict with each other, practical philosophy says, not that the stronger obligation takes precedence, but that the stronger ground of obligation prevails" (*MM* 6:224.) It is easiest to make sense of this (as McCarty does) as a claim that although we can never be obligated to do conflicting actions, and that duties themselves are consistent, the connections between duty and obligation, i.e. the grounds by

which the duty obligates, can conflict. Even granting that it is grounds that are the relevant conflicting objects and that duties themselves "form a morally consistent set" (McCarty 68) this does not resolve practical moral quandaries, since we have no way of knowing what the "stronger ground of obligation" is.

I would like to argue, in fact, that any method of determining the stronger ground of obligation in a situation of moral conflict is essentially a CI-Procedure and falls victim to its same pitfalls. McCarty uses Kant's conceptions of perfect and imperfect duties to begin to develop a theory of strength of grounds of obligation (i.e. perfect duties

give stronger grounds of obligation than imperfect ones.) Setting aside for a moment the difficult question this approach leaves open regarding conflicts between two grounds both generated by (im)perfect duties, we can see that if this were to be a successful approach, we would need to be able to reliably distinguish perfect from imperfect duties. Recall that Kant, in the *Groundwork*, characterized perfect duties as those for which a violating action has a maxim that not only cannot be willed to be universal, but a world in which it is a universal law is inconceivable. Imperfect duties, on the other hand, are those that merely cannot be willed to be universal, but could be conceived of (*G* 4:422.)

Distinguishing between perfect and imperfect duties, then, requires analysis of what maxim appropriately describes a certain action, and analysis of a hypothetical world with that maxim as universal law. Take, for example, a situation in which someone confides a deep, dark secret of theirs in me. They do not want me to share it with anyone else and tell me so. After being told, I get the feeling that I simply must pass the secret on, but only to a single person, my best friend. This satisfies my desire to gossip. According to at least one analysis, then, my maxim is "when entrusted with a secret that the teller wants not to be spread at all, only share it with your single most trusted confidant and otherwise do not pass it on." Let us consider the universalization of this maxim. If everyone were trustworthy when it comes to secrets, except in the case of telling one other person, would the institution of secret-sharing destroy itself? It is unclear, I think. On the one hand, no one can plausibly swear another person to absolute secrecy at all, since like in the classic promise-keeping example, everyone knows that this is a complete pretense. On the other hand, the spread of secrets will be slow, and the classic exponential leak situation where each person tells, say, five more people until the whole town knows will be avoided. It is not clear whether this known, but limited, breach of trust baked into the institution of secret-confiding is an inherent contradiction. If it is, we may say the duty not to tell secrets is perfect. If it is not, we may then say that despite

no inherent contradiction, we cannot will such a world to be since it would predictably eliminate our own ability to confide in others with actual secrecy. Thus, we would say it is an imperfect duty.

We can see in this analysis an exact recurrence of the steps of the CI-Procedure. Just as in the CI-Procedure, we are faced with the difficulties of determining the correct level of specificity of circumstances. In the determination of whether a particular hypothetical world is conceivable or not, there are echoes of Rawls's fourth step that expects one to "calculate as best we can what the order of nature would be once the effects of the newly adjoined law of nature have had a chance to work themselves out" (Rawls 83.) The tools required to distinguish perfect from imperfect duties are very similar to those required to sort out the use of the categorical imperative head on. In other words, if we abandon the CI-Procedure as a correct description of first order Kantian prescriptions, then it does us no good to look instead to the project regarding the strength of the grounds of obligations that Kant sets out in the *Metaphysics of Morals*. It is easy to see that a grounds-procedure with which we can determine which grounds of obligations are stronger than which is easily convertible into a CI-Procedure (and vice versa.)

We can see such problems riddled throughout the *Metaphysics of Morals*. Consider for example a common defense Kantians use to wiggle out of the problem of difficulty of application of the categorical imperative: "If we are to avoid a common misunderstanding, we need to be clear from the beginning that Kant did not hold or teach that we need to appeal to the categorical imperative every time we act or are faced with a difficult decision. The function of the categorical imperative is to help us generate maxims – general rules or policies – not actions" (Sullivan 3.) This maneuver is supported by Kant himself, who in his description of "wide" duties described them as those that "can prescribe only the maxim of actions, not actions themselves" (*MM* 6:390.) It seems, however, that when Kant applies this doctrine, it serves not to provide a

practical framework regarding maxims (as opposed to actions) it instead justifies leaving holes in the practical guidance of the theory. In discussing the limits of the duty of benevolence he says that "how far [the duty of benevolence] should extend depends, in large part, on what each person's true needs are in view of his sensibilities, and it must be left to each to decide this for himself… the duty has in it a latitude for doing more or less, and no specific limits can be assigned to what should be done" (*MM* 6:393.) Indeed, for the reasons discussed above it seems such holes must exist. If Kant had successfully elaborated a procedure for choosing right maxims or for sorting out the priorities of duties, then the tools developed therein would almost certainly be able to provide a CI-Procedure for action. The evasive maneuver of Sullivan is only successful if the retreat goes beyond just claiming that Kantian morality is for guiding maxims not action, instead it must include a large amount of Kant's "latitude."

We return, then, to Wood's version of Kant empty handed of first order prescriptions that directly follow from Kant's theory (or any successful method for generating them.) Is this, then, so big a problem for Kant after all? We have seen that both Kant and his interpreters at least sometimes see this as a beneficial feature of the theory; Wood could claim again that all that is necessary for right action is "common human reason" unsupplemented by any logically pinned-down guidance from the categorical imperative. Certainly, a Kantian could concede the point that the theory is more or less empty of first order prescriptions, and instead see the project as solely metaethical in nature. In this way, all complaints about a moral procedure would be ill-founded, since in this interpretation the domain of Kantian ethics is merely to explicate the source of obligation. One could even take this tack while still retaining some normative content, such as for example the claim that the central objects of morality are duties. Perhaps this metaphilosophically explains some of the work of contemporary Kant scholars, such as Thomas E. Hill, who does much analysis on the work of Kant, but in his

practical philosophy stresses that his discussions are "often Kantian in spirit, but there is no attempt here to do textual exegesis or to crank out solutions from Kant's theory" (Hill 1.) My criticisms of Kant are irrelevant if Kantian practical ethics is done this way, with only a vague, elliptical Kantian spirit that morality is the stuff of duty without self-exception.

However, I do not think this is how Kant (or Wood, really) sees his own philosophy. The *Metaphysics of Morals* contains many instances of "casuistical questions" involving particular scenarios (in part as a sort of exercise to the reader.) If Kant's philosophy is not supposed to give a binding answer to these questions, it is implied that it is at least supposed to be a strong guide. Insofar as these exercises are merely meant to engage and sharpen the faculty of judgement in the way that Wood stresses, it seems then that the answers are coming only from moral intuition, rather than any philosophy at all. Again, a theory of ethics that combines a Kantian source of obligation with a sketchier practical system based on intuitions does not have the problems discussed here, but at the same time this would mean Kant should have nothing to say about hard ethical cases.

Let us look at what Kant had to say about one hard ethical case in particular, the infamous case of lying to a murderer, in order to see in what way his categorical imperative is applied. In the essay "On a Supposed Right to Lie From Philanthropy," Kant argues against the utilitarian philosopher Benjamin Constant that our duty to tell the truth extends even to a situation in which a murderer shows up at our door asking for the location of a possible victim. Kant argues that in this case the duty to truthfulness supersedes any duty of general beneficence, i.e. that "every individual… has the strictest duty to truthfulness in statements that he cannot avoid, though they may harm himself or others" (*SRL* 8:428.) The form of his argument seems to be, at its bare bones, that since the duty to truthfulness is a perfect and unconditional duty (an argument for which would

likely go similar to the one presented for the specific case of a duty not to break promises about loans given at *G* 4:423) there is no exception even in the case of lying to a murderer. In particular, this logic is used to reject Constant's claim that "To tell the truth is a duty, but only to one who has a right to the truth." Kant never considers in the essay possible other applications of the categorical imperative. In particular, he does very little analysis of the duty of beneficence, even though his conclusion implies a resolution in this case in a conflict between that duty and a duty of truthfulness. Instead, his arguments consist of other kinds of reasoning, ones in which it is emphasized that if a lie is told and the murder happens anyway, then the liar is responsible for the harm, but that if it happens after the truth is hold then "an accident *causes* the harm" (*SRL* 8:428.) Instead of explaining the reason why the duty of truthfulness is inviolable in this case where it seems difficult to select the appropriate circumstances for a maxim and where there is a conflicting duty of beneficence, Kant merely reiterates the dogma that duty is about not asking for self-exceptions. He criticizes someone who even thinks of lying by saying that someone who "asks permission to think about possible exceptions [to the duty of truthfulness] is already a liar" (*SRL* 8:430.)

Of course, this essay is controversial even among Kantians. Michael Cholbi suggests that lying to the murderer is actually required under the Kantian conception of self-defense. Christine Korsgaard thinks that the lie is permissible under the Formula of the Law of Nature (the one on which we have so far focused) but not under the Formula of Humanity. It is not, then, Kant's own implausible claim that we must not lie to the murderer that really gets at the heart of the problem here, as has been often thought. Instead, the problem is that such a wide variety of disagreement among Kantians is even possible. Is the misunderstanding of Kant's theory that widespread, extending even to the man himself? Or, more likely, are all these casuistical answers compatible with Kant's theory? If the latter, then it follows that the practical content of Kantianism is vastly

underdetermined. Uncharitably, this makes it seem like sophistry when particular answers to moral quandaries are derived from Kantian duties, since the implication is that such an answer is *the* answer that follows from a correct application of the categorical imperative to this case, when in fact many such answers are possible (even, perhaps, consequentialist-type answers that almost no actual Kantian would endorse.) Any Woodian hope that decision procedures for hard problems are unnecessary and that in fact all that is needed is a refined sense of judgment, an appreciation for the concept of duty, and an unwillingness to make exceptions for oneself is shattered when we look at how strongly a group of self-professed Kantians can disagree with each other about casuistics.

In sum, I'd like to suggest that Kantian ethics is caught between a rock and a hard place. If it is put into a form with obvious practical content, such as with a CI-Procedure, then it becomes subject to problems like puzzle maxims and being divorced from its metaethical bedrock. If, on the other hand, the purely metaethical nature of the theory is embraced, then its practical content withers to nothing, unable to give guidance on which maxims or which duties take priority (since doing so would be eventually equivalent to a CI-Procedure.) The middle road that many Kantians, including Kant himself, take wherein the gaps in the practical theory are filled in by intuition, all the while falling back on the unrelated metaethical component of the theory as justification, is untenable. This is what makes it possible for Kantians to have such a wide variety of incompatible opinions on practical ethics, all of which are supposedly grounded in the categorical imperative. A more reasonable moderate strategy for deontological ethics might be like the aforementioned one Hill's work takes, in which applications of normativity are heavily flavored by Kant's conception of duty, but the pretense is dropped that all of practical ethics consists ultimately of special cases of the categorical imperative.

Works Cited

Bittner, Rudiger. “Maximen”, in *Akten des 4. Internationalen Kant-Kongresses*, edited by Gerhard Funke, Bouvier 1991.

Brentano, Franz. *The Origin of Our Knowledge of Right and Wrong*. Routledge, 1969.

Cholbi, Michael. “The Murderer at the Door: What Kant Should Have Said.” *Philosophy and Phenomenological Research* 79, no. 1 (2009): 17–46.

Hill, Thomas E. *Autonomy and Self-Respect*. Cambridge Univ. Press, 2000.

Kant, Immanuel. *Practical Philosophy*. Translated by Mary J. Gregor, Cambridge University Press, 1999.

Korsgaard, Christine M. “The Right to Lie: Kant on Dealing with Evil.” *Philosophy & Public Affairs* 15, no. 4 (1986): 325–49.

McCarty, Richard. “Moral Conflicts in Kantian Ethics.” *History of Philosophy Quarterly* 8, no. 1 (1991): 65–79.

Mill, John S. *Utilitarianism*. The Floating Press, 2009.

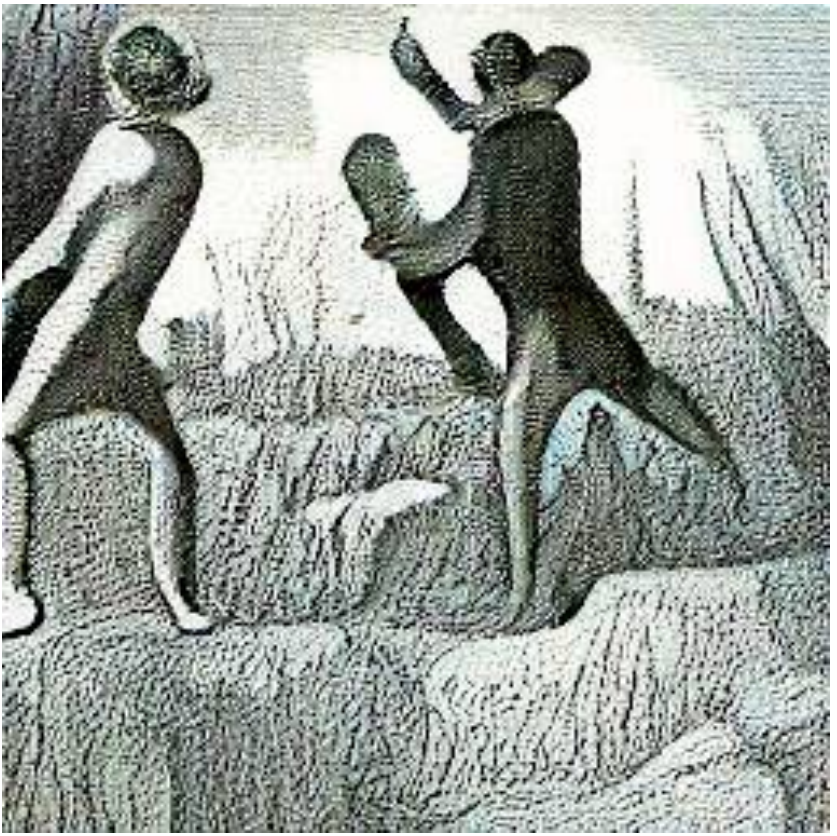
Rawls, John. “Themes in Kant’s Moral Philosophy” in *Kant’s Transcendental Deductions* edited by Eckert Forster, Stanford Univ. Press 1989.

Sullivan, Roger J. Introduction to *The Metaphysics of Morals* edited and translated by Mary J. Gregor, Cambridge University Press, 1996.

Timmermann, Jens. “Appendix C: Universal legislation, ends and puzzle maxims” in *Kant’s Groundwork of the Metaphysics of Morals: A Commentary*. Cambridge Univ. Press, 2007

Wood, Allen. *Formulations of the Moral Law*. Cambridge Univ. Press, 2017.

Pulling for Moralism: Rough Heroes and the Moral Aufheben Argument



David Veldran

I. Introduction

In "Robust Immoralism," Ann Eaton (2012) introduces the rough hero, a character we morally disapprove of, but one for whom we have sympathy, affection, or admiration. Positing that moral flaws in works of art can be aesthetic merits, Eaton argues that some rough hero works (RHWs), to the extent they endorse an immoral character, are morally defective and aesthetically good therefore. Noel Carroll and other "moralists," who generally hold a tighter relationship between the morally and aesthetically good, resist Eaton's claims. It is true that Carroll's (1996) "moderate moralism,"[73] which I will focus on, is not in direct conflict with Eaton's immoralism, but the two views are currently locked in a tug-of-war, wrangling over specific cases.[74]

In this paper, I respond to Eaton's arguments for immoralism and support Carroll's moderate moralism. I analyze several works, mostly films, to show that many seemingly immoral works are in fact moral, though in a way many moralists, including Carroll, have overlooked. While I agree with Eaton that RHWs challenge our moral intuitions by prescribing admiration for immoral characters and evoking "delicious" ambivalence (an aesthetic merit), I don't find this challenge, or the works, *eo ipso* immoral. On the contrary, I argue, it often serves morality by helping to *improve* our moral intuitions. The paper has roughly two parts: first, I outline three rebuttals to Eaton's immoralism and show why a fourth—my moral *aufheben* argument—is necessary. Then I show how my view allows (moderate) moralism to absorb Eaton's most challenging RHWs.

---

[73] Moderate moralism holds that "some works of art may be evaluated morally…and that sometimes the moral defects and/or merits of a work may figure in the aesthetic evaluation of the work." (p. 236).

[74] While Carroll generally thinks moral merits will be aesthetic merits, and moral defects aesthetic defects, he doesn't rule out the possibility of switch-hitting. However, he says he has never seen a compelling example of immoralism (2013, p. 371).

II. Three Rebuttals to Immoralism

Eaton's (2012) RHWs are supposed to be examples in which moral defects in a work are aesthetic merits. However, like Carroll, I am skeptical that the elements Eaton identifies —in brief, prescriptions of admiration for immoral characters—are in fact moral defects. Here are three rebuttals to Eaton's claim. One is inspired by Aristotle (2016) whose *Poetics* suggests that tragedy may serve morality by evoking and purging [catharthis] pity and fear. Similarly, one can argue that an apparent moral defect in a work serves morality through its purgative or purifying effect—and so is not truly defective. Living vicariously through the likes of Travis Bickle and William Munny, perhaps we, in a controlled environment, exercise—and so exorcise—our immoral impulses, helping to redeem the work morally. A second rebuttal is inspired by Jacobson (1997) and Kieran (2003), who —perhaps unwittingly[75]—offer another way in which a moral defect can serve morality. By showing how others are in error (Jacobson) or allowing us to "[experience] what's bad to understand the good" (Kieran, p. 63), a moral defect may, one could argue, lose its defectiveness.

These arguments are unsatisfactory, however, because the way these moral defects serve morality might be entirely extrinsic to the work. That is, the moral lessons the work teaches might in fact be taught by others (audiences) who use the work as a kind of prop. Similarly, Dadlez's (2017) objection to immoralism—that, since the "moral confusion" RHWs produce is unlikely to change our moral beliefs, there's no moral defect—fails to chip away at *intrinsic* immorality. The advantage of the third rebuttal, Carroll's (2013) narrative argument, as I'll call it, is that it dissolves works' alleged immoral elements into mere depictions—rather than endorsements—of immorality, making the moral lessons intrinsic to the works. According to Carroll, we should consider a (narrative) work's apparent immoral elements in the context of the work *in toto*. If the

[75] These arguments are meant to defend *immoralism*, but as Eaton (2012) points out, they collapse into moralism.

narrative condemns the immorality it depicts, the element (the depiction, not the depicted) is not immoral, for it teaches us moral lessons. This accords with Hume's original quote,[76] where we find the qualification "without being marked with the proper characters of blame and disapprobation," suggesting that all is well if the immorality is condemned.

III. Two Kinds of Rough Heroes

Though it has its limits, the narrative argument can win back several rough heroes for moralism. I will call these figures the Martyrs, those rough heroes who receive their just deserts and so figure in a work's overall moral message. I regard characters like Darth Vader, Norman Osborn (Green Goblin), and *Breaking Bad's* Gus Fring in the same way that a Yankee fan hates, but does not despise, the Red Sox. Just as a Yankee fan needs the Red Sox in order to relish his own team beating them, perhaps my affinity for these characters is but a sign of my (moral) desire that good triumph over evil. Here, even my admiration for them isn't necessarily immoral: admiration and hatred may be compatible, for I can root for the Yankees, hate the Red Sox, and admire both teams' success without committing treason. Similarly, I can admire these hateful villains and give devils their due without endorsing them. My admiration merely acknowledges them as worthy opponents.

Come to think of it, there are several rough heroes on Eaton's list I don't admire. I'll call them the Spiders—those who, if we don't completely despise them, evoke a spidery disgust, despite their positive portrayal. Figures like Humbert Humbert, Hannibal Lecter, Alex from *A Clockwork Orange*, and *American Psycho*'s Patrick Bateman are fascinating, but not, I think, admirable. Eaton thinks some are, but for

[76] The quote cited by Eaton (2012) and others in this debate begins, "where vicious manners are described, without being marked with the proper characters of blame and disapprobation; this must be allowed to disfigure the poem and be a real deformity…" (Hume, 1987).

myself, I'd rather *watch* them than associate with them or *be* them, if only for a day,[77] and I certainly don't think I have anything to learn from them. I regard these ignoble figures —more creature than character—as I regard *Triumph of the Will's* Hitler, who, while despicable, is hardly uninteresting. But one's *interest* in psychopaths or mass-murderers isn't immoral—one's *admiration* is, and none is available for Spiders.

Moreover, the affinity I have for them is not primarily due to their likable qualities, as Eaton contends. I detect in myself a similar fondness for the Wicked Witch of the West, Gollum,[78] and Freddy Kreuger. These characters aren't rough heroes—they lack the humanizing portrayal Eaton describes—and yet, I unabashedly relish the scenes in which they appear. My affinity for them is borne of my own curiosity about evil, not of a moral defect in the work.

IV. The Moral Aufheben Argument

The narrative argument can only get us so far: I think Eaton (2013) successfully shows that it meets its end at the hands of Tony Soprano. For Eaton, *The Sopranos* is morally defective (therefore aesthetically better) to the extent it endorses its immoral protagonist (2012, p. 282). Carroll, on the other hand, views the show's treatment of Tony as didactic, asserting that it may warn us "not to allow our moral radar to be jammed by…irrelevant moral static," like Tony's wittiness (2013, p. 372). In response, Eaton (2013) insists that audiences cannot keep their nonmoral approval from contaminating their moral disapproval, and I agree. It is exceedingly difficult to parse out the good and bad in Tony, and I find myself in limbo with him, which is quite "delicious." Nor, as Eaton points out, does anything in the show directly rebuke my admiration for the man; Carroll's narrative Authority is but absent. The question arises: have we reached the limits of moralism, or might another argument account for Tony Soprano?

---

[77] I'd happily have a beer with Tony Soprano and wouldn't mind being him for a day, if that's a useful barometer.

[78] Not only is he hideous, he's not even smart (he's bad at riddles). I'm referring to his depiction in the first two books/movies, before he redeems himself morally.

To account for Tony, I'd like to build on a part of Stecker's (2008) attack on immoralism. According to Stecker, whether a work contains a moral flaw does not supervene upon whether our prior moral intuitions match those the work exhibits. Just as a match might add nothing of moral value to the work because the intuitions are so banal, a mismatch might be a moral merit if it offers us an alternative "reasonable moral assessment of [a situation]" (p. 158). What seems to be a moral flaw might be a forgivable "error in judgement," and so we may praise an allegedly immoral work (and call it moral) "for exploring an alternative that has some claim to be true in its own right" (p. 159).

To apply Stecker's insights to RHWs, I must make two modifications. First, RHWs *endorse* (not only "explore) characters with immoral perspectives, as Eaton (2012) shows us, by prescribing admiration for them. Second, we should apply "reasonable moral assessments" cautiously: many rough heroes behave unreasonably, and their hamartia is often worse than a mere "error in judgement." But for other rough heroes, these terms are not far off the mark. For instance, Tony's judgement that crime is the best way to support his family is erroneous, but not entirely unreasonable—not, at least, in the way that sponsoring gratuitous torture (as Spiders often do) is. Let's compromise that rough heroes like Tony have immoral but "*sort of* reasonable" intuitions. Even with these qualifications, RHWs can still (intrinsically) serve morality, so I argue.

With Stecker, I contend that we, a work's target audience, may have flawed or incomplete moral intuitions that seemingly immoral works can rectify. Going beyond Stecker, I think a work that *endorses* characters' "*sort of* reasonable" intuitions can be moral: if these intuitions have "some claim to be true in [their] own right," the work, by espousing their owners, may have something to teach us about morality. Some RHWs do this through what I call moral *aufheben*.[79] By challenging their target audiences' flawed

[79] The name is after Eaton's invocation of the word to describe how rough heroes overcome our imaginative resistance (2012, p. 287).

moral intuitions and offering them new moral truths (or *sort-of* truths), a work may, I submit, serve morality. And serve it intrinsically—the moral lessons are its own. Importantly, I think they do not merely destroy their audiences' prior intuitions (which also have some claim to truth) but, in Hegelian fashion, at once cancel, preserve, and lift them up,[80] thereby improving them. In this way, a work may intrinsically help us, following Eaton (2012, p. 288), solve a problem worth solving. A *moral* one, I hasten to add.

V. Why Aufheben?

Except Stecker's and mine, the above arguments for moralism take a limited view of what it means to serve morality. For them, it roughly means to serve our considered views, to use Eaton's phrase. Carroll (1998), for instance, believes that morally good works can improve our moral intuitions, but not so much by challenging them: his "clarificationist" view holds that moral works "deepen our moral understanding" not by giving us new moral knowledge, but by teaching how to apply our present knowledge (p. 142). Let's canvass this trouble with this view. Carroll's leading example of a *clarifying* work is *A Raisin in the Sun*, which, he argues, allows white audiences to *understand* what they already *know*, that African Americans are people and deserve equal treatment. Per Carroll, the work encourages audiences to apply this knowledge when, in the play, a black family encounters discrimination.

But if this were simply a case of being "prompted to make connections between the beliefs [white audiences] already possess" (p. 143)—rather than gaining new moral knowledge—why should the play go to lengths to humanize African Americans as Carroll says?[81] If white audiences already knew they were persons, this element would be extraneous, even distracting. We can read the play now as if we knew that blacks were

[80] See Kaufmann (1974. P. 236) for these three meanings of *aufheben*.

[81] Carroll says the play shows "that the dreams and the family bonds of the major characters are no different from those of other persons" (p. 143).

persons, but the work doesn't assume we know this; if we do not *understand* what we know, it asks Carroll, do we really know it? Contra Carroll, I suggest it is *Raisin's* challenging the audience's prior (racist) beliefs—not affirming and teaching how to apply them—that gives it its moral bite. Here is what I suspect occurs when a *Raisin's* target audience receives the work: the audience comes in with flawed moral priors, those priors are challenged, and the audience leaves with better moral intuitions—not perfect ones, nor necessarily the self-same ones espoused by the work, but better ones.[82]

Challenging our moral intuitions is not a rare way to teach moral lessons. Consider how such clearly moral films as *Crash* and *In Bruges* challenge our prejudice that immorality is for other people—specifically, for cold-blooded monsters. Encouraging us to admire the wrongdoer, just as Dostoevsky has us admire a wayward but all-too-human Raskolnikov, these films teach that everyone is an amalgam of good and bad. *Bruges'* Ken is not only a hitman: he is an honest and loyal friend; nor is Ray just a (accidental) child-killer—as the film stresses, he can redeem himself; and even the villainous Harry has admirable integrity, as he demonstrates by his "you've got to stick to your principles" suicide. *Bruges'* moral lesson, if agreeable in abstract, is unsettling in the moment. Like *Crash*, which portrays, inter alia, a racist police officer who rescues a woman he once assaulted, *Bruges* challenges our moral priors, but it is for our benefit.

These challenges don't make us doubt everything, but neither do they simply affirm what we already know and show us how to apply it. When *Do the Right Thing's* Mookie hurls a trash can at Sal's pizzeria, igniting pandemonium, the film directly challenges our bias against violent civil disobedience. As the film's competing epitaphs—quotes from Dr. Martin Luther King Jr. and Malcolm X—emphasize, the film offers new moral knowledge: that violence may be an appropriate response to racism. We need not have already believed this to find Mookie's action just—the film can generate this belief

[82] I further suspect that this benefit is renewable and additive, such that each new interaction with the work can continue to benefit the same audience.

on its own, just as *Raisin* and *Crash* can alone convince racists of their error. That many will initially recoil, as I did, from Mookie's action—or from the invocation of Malcolm X, often regarded as an extremist—does not indicate a moral defect. Rather, if *Do the Right Thing* is (even partially) right about violence, this jolt may be just what its target audience needs to improve its views via *aufheben*.

VI. Admiration, Endorsement, and Moral Teachers

Let's shift gears to consider the morality of admiration and endorsement, a topic that will help us understand how RHWs like *Sopranos* are moral. Here, I'll use admiration and endorsement interchangeably as I think that (a) admiration for an object is an effective endorsement of it and (b) the way works endorse is by prescribing admiration.[83] Our admiration for characters in the above "clearly moral" films is justified, I think, not only because they are moral, but because they teach us moral lessons. Of course, these often go together (Mookie), but some characters, like *Bruges*' Harry and Ray and *Crash's* Anthony, Jean, and Farhad,[84] teach us moral lessons without themselves being moral. Nor are their lessons merely the narrative's condemnation of them: Harry's suicide, for instance, is a lesson in integrity that *he* teaches—it's not simply a Carrollian narrative condemnation of evil. Harry is thus a member of a third class of rough heroes: the Teachers.

This brings me to a larger point, that admiration for an immoral character is not necessarily immoral. When we admire someone, I suggest we are holding him up as a kind of good teacher, broadly construed. That is, we are acknowledging that he may help us solve a problem worth solving. Admiring someone with virtually nothing worthwhile

---

[83] I focus on admiration as I think Eaton's other children, sympathy and affection, are not as strongly tied to endorsement (I feel some sympathy for Kreuger and some affection for Humbert, but approximately zero *admiration* for either).

[84] Anthony is an unprincipled carjacker, whose only moral merit is that he frees twenty Asian slaves he inherited, (I think another Anthony, the mob boss, would have done the same). Similarly, Jean doesn't make a dent in her bigotry by hugging her Hispanic housekeeper, and Farhad, who ragingly tries to kill an innocent locksmith, is more stunned than self-reflective when his gun fails.

to teach us, like Riefenstahl's Hitler, is immoral. But it is simply not true that all immoral characters and rough heroes have nothing worthwhile to teach us. Still, it's not compelling that the fact that a rough hero has *something* worthwhile to teach us warrants admiration for him (should we admire Hannibal Lecter, say, if he can teach us how to juggle?). Thus, the following argument is insufficient:

1. Admiration is morally warranted for good teachers.
2. Good teachers help us solve a worthwhile problem.
3. Some rough heroes help us solve a worthwhile problem.
4. These rough heroes are good teachers.

C. Admiration for them is morally warranted (Teacher-RHWs are moral).

Admiration for the specific thing they teach is warranted, but if Eaton is right that we cannot parse apart rough heroes' good and bad characteristics, then admiration for them —for their *whole character*—seems as immoral as they are. Thus, Eaton may counter that I need to factor in the following and conclude that admiration is not warranted:

5. Admiration is anti-warranted for immorality.
6. Rough heroes' immorality outweighs their good pedagogy.

As I grant (5), I intend to overturn (6) to justify admiration for Teachers (and so call Teacher-RHWs moral). Some rough heroes' good pedagogy outweighs their immorality, I argue, because their immorality uniquely—and intrinsically—positions them to help us solve certain worthwhile problems via *aufheben*.[85] Whereas juggling can be taught just as well by a saint as a sinner, some worthwhile lessons, I maintain, cannot —and they are worth the extra immorality required to learn them.[86] Thus, I propose that C holds because some rough heroes (Teachers) meet the following:

---

[85] I recognize that these might be (or are) extrinsic facts about the moral lessons, meaning that RHWs are not themselves *intrinsically moral*. But I didn't claim that: all I said was their *lessons are intrinsic to them* and so, by teaching them, *they intrinsically serve morality*. To justify this whole enterprise as moral, I'm happy to rely on some extrinsic facts, which I defend below.

[86] To be so worth it, they de facto will always be *moral* lessons (Super-juggling, which only Hannibal can teach, say, is plausibly not worth our admiration for him, given what else he represents).

6A. One's good pedagogy outweighs their immorality iff his immorality uniquely and intrinsically allows him to teach valuable moral lessons via *aufheben*.

Interestingly, this pits me against the bulk of Stecker's (2008) argument in the very article I build upon: Stecker and I agree that seemingly immoral works may teach us moral lessons, but he thinks that doing so is not intrinsically tied to *endorsing* immorality. In dissent, I offer Tony Soprano:

VII. Learning Morality from a Gangster

Consider what lessons Tony can teach us that less immoral characters cannot (or teach far less effectively). First, from Tony we can learn that most people, even gangsters, are morally complex. (*Unforgiven* and *Pulp Fiction,* two RHWs Eaton cites, teach similar lessons about cowboys and hitmen, respectively). Eaton (2012) acknowledges that rough heroes are not bereft of moral virtues, but she really ought to give Tony more credit. When we admire him, we endorse someone who does not only have moral flaws and non-moral merits, as Eaton and Carroll all but suggest, but someone with pluses and minuses in *both* categories—moral and non-moral. Tony is immoral, no doubt, but he's also "sort of reasonable:" he's principled, devoted to family, often honest, and even, at times, merciful.[87] He thus cautions me against neatly dividing people up into categories of good and evil, inviting me to appreciate their nuances. *Crash* and *Bruges* teach this too, but to the degree they don't endorse immoral characters as much as *Sopranos* does—and so challenge our moral priors as strongly—I think they are less effective.

Similarly, the Eatonesque ambivalence Tony generates in me is perhaps itself a moral lesson: at once admiring and recoiling from Tony, we may learn that morality is not as simple as we often assume, that good and evil are more like day and night—lacking a

[87] Eaton and Carroll seem to miss Tony's many moral virtues: in addition to trying to be a good father, he reconciles his misdeeds with his attempt to support his family; he is loyal to his friends; he values honor and respect; he has a moral compass and is no psychopath; he is often honest with himself and others (he even goes to therapy!); he feels guilt and remorse—for instance, when his cousin returns from prison for a crime Tony was supposed to commit; and he is, of all things, merciful: he doesn't celebrate Vito's homosexuality, for example, but he also doesn't think it warrants death (a minority view).

clear dividing line—than left and right. Even if Tony is immoral on balance, his flaws, as Eaton points out, are so entangled with his virtues that his portrayal complicates our moral judgments and "mudd[ies] the waters" (2013, p. 376-7). But I don't think this muddiness signals a moral defect—rather, it may be its own moral lesson, another way of showing us, with *Talk to Her's* haunting epitaph, that "Nothing is simple."[88] If this is indeed a worthwhile lesson, I can think of no better way to teach it than by getting an audience to admire an immoral, but "sort of reasonable" character.

Most notably, Tony teaches the value of devotion to family. By portraying devotion as if it were far more important than others' suffering (I am thinking of Eaton's (2013, p. 377) "curb stomp" example, an immoral but "sort of reasonable" way to protect his daughter), *Sopranos,* via Tony, sears its moral value into my mind. Consider that the show's trademark scene is not Tony ranting about enemies or engaging in crime, but him toasting to his family:

> *To my family. Someday soon, you're going to have families of your own, and if you're lucky, you'll remember the little moments, like this…that were good* (S1, E13).

That a show about gangsters can revolve around this scene—and make me feel its moral pull as if I were a child at Tony's table—is a testament to both aesthetic and moral merit. Tony, I submit, does not only help solve Eaton's aesthetic problem of delicious ambivalence; he also helps me solve moral problems: Tony shows me what counts in life, and his violent, dramatic, and only "sort of reasonable" way of doing so is, I think, necessary. Who can better teach the value of devotion to family than men like Tony, Walter White, and *Mystic River's* Jimmy Markum, who would sooner turn the world upside-down than rest at a harm done to their family?

---

[88] Tony's moral entanglement may not evince Carrollian (2013) "moral clarity," but why should clarity be our standard? If morality *itself* is not clear, clarity can only teach us so much.

I emphasize *help* me solve problems; my ambivalence towards them keeps me from uncritically accepting Teachers' lessons and letting them solve them for me (only bad teachers to that). Here is how I envision the *aufheben*: I initially think I believe in devotion to family, but I really don't. In Tony, I encounter beliefs that challenge mine, but perhaps believe in devotion too much. Both my and Tony's beliefs have "some claim" on truth, and our values are part moral and part immoral. When they collide, I think their truth is preserved and some of their falsity is cancelled, leaving me with a better set of beliefs than I had before. Thus, from Tony I do not learn that it is right to maim those who bad-mouth my family—I just learn that, as family is truly what counts in this life, I should up my devotion.

This is new moral knowledge: despite what I may have advertised to myself and others, I simply did not (functionally)[89] know that family was what counts. In learning this, however, I don't simply abrogate my moral priors and, as Eaton says, turn "against the forces of good" (2012, p. 285). In admiring Tony, rather, I participate in a kind of dialogue with an immoral, but not despicable, opponent, and it is for the benefit of (my) moral improvement. There is collision, but no collusion.

VIII. The Immorality is the Point

Perhaps Eaton will object that I overlook that RHWs themselves regard their heroes as immoral and that we don't understand the work unless we appreciate their immorality (2012, p.283). How, then, can they teach us moral lessons? Because, I submit, teachers' immorality is inseparable from their moral lessons. Thus, to gain these lessons, we must appreciate their immorality. To understand just how important devotion to family is, we must appreciate that characters like Tony are willing to do deeply immoral things for their sake (if they drew the line at immorality, they might fail to convince us they *really*

[89] I might think I know, but my actions tell a different story.

cared). Thus, even admiring a teacher *for* his immorality, as Eaton believes we sometimes do, is justifiable if his immorality is intrinsic to his good pedagogy.

Films like *Fight Club* and *Talk to Her* expertly blend teachers' immorality with their pedagogy. Tyler Durden, for instance, teaches us to live authentically through his violent anarchism, not despite it. Sans Durden's hellraising, there is no moral lesson—or it is far weaker. When he pulls a gun on a shop owner, threatening to kill him unless he snaps out of his Sartrean bad faith, he teaches existentialism more effectively than such peaceful figures as Kazantzakis' Zorba and Hesse's Siddhartha. Like God asking Abraham to prove his faith not with words but with his son's blood, Durden raises the stakes of his lesson to the point of immorality, thereby driving it home.

Similarly, *Talk to Her*'s Benigno teaches us (improbably) how to care for others through—not despite—his rape of Alicia. The film's sympathetic treatment[90] of his act does not, contra Eaton (2008, p. 18), apologize for rape. Rather, it serves a broader moral purpose—prescribing admiration for a character who, while deeply flawed, teaches Marco and us, as Shpall (2013) argues, how to care for and "talk to" others and treat them as ends-in-themselves. Benigno's rape of Alicia is not, paradoxically, a selfish act, nor one that treats the latter as mere means: in Benigno's mind, it's a mutual, even consensual one, a natural part of their relationship. If Tony's "curb stomp" takes devotion to family too far, Benigno's rape takes his selflessness to its perverse extreme. Both, thus, teach us moral lessons via their immoral, but "sort of reasonable," intuitions.

Because the sympathetic treatment of the rape serves morality, *Talk to Her* does not apologize for rape *simpliciter* (if it did, it would be immoral). That is, the extent to which it does apologize for rape is instrumental, and necessary, to the moral lessons it teaches. In the same way that *Talk to Her* is not fundamentally (as Eaton seems to imply

---

[90] Not only does the film encourage us to pity the naïve Benigno, but it romanticizes his rape of Alicia as Eaton (2008, p. 18) says, even anticipating it with *The Shrinking Lover*, a silent which, as Eaton points out, winks at nonconsensual penetration. Amparo, the shrinking man's lover, enjoys the experience, and Benigno wakes Alicia from her coma by raping her, two points that, along with the dance teacher's "Nothing is simple" remark, give us mixed messages about Benigno's transgression.

it is) a meditation on whether it's permissible for a childlike nurse to penetrate his unconscious patient, *Pulp Fiction* is not, as Carroll (1996, p. 230) seems to think, in the business of judging that rape is worse than murder. That one scene in the latter film *might* suggest this is not *eo ipso* a moral defect. On the contrary, the scene Carroll refers to serves morality by teaching us to love our enemies.[91]

In admiring immoral characters like Durden and Benigno—and Jules and Butch for that matter (Vincent is a Martyr)—we attest that they, with their "sort of reasonable" moral intuitions, have something to teach us. Perhaps, like *Fight Club's* Narrator, we are deficient in authenticity, or, like Marco, in empathy—and we need a teacher, even an immoral one, to set us right. Note that neither Durden's nor Benigno's death functions merely to condemn their immorality. Rather, it seals their moral lessons, turning them into kinds of martyrs. In killing Durden, the Narrator shows that he's *learned from him* (indeed, his new authenticity allows him to be with Marla). And in losing Benigno, Marco can finally absorb what his friend taught about connecting with others, allowing him, of all things, to be with Alicia.

Ridding themselves of the excesses of their foils, Marco and the Narrator can still incorporate what was valuable and "reasonable" in them, completing an *aufheben* that leaves them more moral. Like the Narrator's and Marco's, when our moral intuitions are challenged by Durden and Benigno, they do not perish—neither we, the Narrator, or Marco suddenly believe that the evils of society justify anarchy, or that devotion to women justifies rape. Rather, I think they get *aufgehoben*, becoming more nuanced—and true.

---

[91] In a film rife with redemptive themes, it's hardly accidental that Butch, after rescuing Marcellus from rape, meets a "chopper" named Grace.

IX. Immoralism's Last Stand

There is one more class of rough heroes I want to address. A subset of Teachers, the Stylists, as I'll call them, may escape the moralist's grasp (though I have my doubts). Taking their cue from Nietzsche, especially *The Gay Science*: 290 and *The Birth of Tragedy*: 5,[92] these rough heroes, in Nehamas' (1985) reading of Nietzsche, make their life into an (literary) art. With Nietzsche, they teach the convergence of aesthetic and moral values (or the former's swallowing of the latter)[93] through their stylistic "immoralism."[94] Some of the above Teachers—Tony, Durden, even Benigno—teach style, but they don't *exclusively* teach it. And there are untold numbers of failed Stylists (Spiders, like Patrick Bateman), who don't even get us to *consider* that aesthetic values could justify their immorality.

A select few—perhaps Nietzsche's Zarathustra, *Rope's* Brandon (accidentally),[95] *There Will Be Blood*'s Daniel Plainview, and *A History of Violence*'s Tom Stall—primarily teach lessons in style, and I think successfully. I find myself admiring them mostly, if not all, for their style. Is admiration for these rogue characters moral? Here, two paths diverge: whereas Eaton and Carroll likely regard style as "irrelevant moral static," Nietzsche insists upon it lest we become dissatisfied with ourselves—"For the sight of what is ugly makes one bad and gloomy" (1974, p. 233). Without engaging in metaethics, I wish to suggest the possibility that Nietzsche's lesson, that art (style) has

---

[92] *Gay Science* 290: "One thing is needful.—To "'give style" to one's character—a great and rare art! It is practiced by those who survey all the strengths and weaknesses of their nature and then fit them into an artistic plan until every one of them appears as art and reason and even weaknesses delight the eye…" (1974, p. 232). *Birth of Tragedy* 5: "…only as an *aesthetic phenomenon* is existence and the world eternally *justified*" (2016, p. 50).

[93] See Came (2014) for Nietzsche's penchant for aesthetic values and his attempts to "replace" traditional morality.

[94] Nietzsche means "immoral" in contrast to Judeo-Christian morality, meaning it is noble and, in a wider sense, moral. See Kaufmann's footnote of Nietzsche's "All experiences are moral experiences…" (Kaufmann, 1974, p. 174).

[95] Brandon, who makes explicit reference to Nietzsche's teachings, is technically a Martyr, but Rupert's closing speech is so corny that it fails to undo my admiration for Brandon as a Teacher.

something to teach us about morality (in the wider, Greek sense of *ethikos*[96]), is worth learning. If style is *not* a problem worth solving (or not worth the immorality required to solve it), admiring Stylists would be like admiring Patrick Bateman for his looks or Hannibal for juggling, cases of immoralism. The Stylists are tricky and require far more analysis than I can provide here, but I'm not convinced a moralist must concede them.

X. Conclusion

Let me briefly summarize. Roughly, the Aristotle- and Kieran-Jacobson- inspired arguments assert that a work that endorses an immoral character can extrinsically serve morality. Carroll's narrative argument holds that a work can intrinsically serve morality if it condemns an immoral character's immorality. My moral *aufheben* argument contends that a work can intrinsically serve morality if it *endorses* an immoral character for moral —or "sort of" moral—ends. It also holds that prescribing admiration for an immoral character is moral if the character uniquely teaches us valuable moral lessons via *aufheben*. The narrative argument can account for rough heroes whose immorality is condemned (Martyrs), and some rough heroes, since they aren't admirable, don't need accounting for (Spiders). The moral *aufheben* argument, I hold, can account for the other rough heroes (Teachers), possibly excepting Stylists.

If my argument succeeds, where does it leave immoralism? Near banished from narrative art? Perhaps Stecker (2008) and other moralists can show that when other forms of art, like humor, broach immorality, they merely *explore* rather than endorse immoral perspectives or they fail to merit the intended response (a joke isn't funny, say).[97] But, as Stecker himself notes, exploration and endorsement are often hard to pull apart: with jokes and, perhaps, visual art, I find myself more susceptible to being pulled from my moral moorings towards unreasonableness. But there is something about stories that

---

[96] Brobjer (2003) argues that Nietzsche, aware of this word's etymology (it comes from *ethos*, meaning character), incorporated it into his (virtue ethicist) morality.

[97] See Gaut's (2007, Chapter 10) "merited response argument," especially as it applies to humor.

keeps me morally awake: many RHWs make me consider how my own beliefs have no monopoly on truth, how I can learn something deeper from seemingly vicious heroes—without sacrificing everything I know. Contra Eaton, some of these heroes—the Teachers—do not only give me pleasure: they teach me how to live.

Works Cited

Aristotle. (2008). *Poetics*. Trans. S.H. Butcher. Project Gutenberg EBook.

Brobjer, T.H. (2003). Nietzsche's affirmative morality: An ethics of virtue. *The Journal of Nietzsche Studies 26*, 64-78.

Came, D. (2014). Nietzsche on the aesthetics of character and virtue. In *Nietzsche on art and life.* Oxford Scholarship Online.

Carroll, N. (1996) Moderate moralism. *British Journal of Aesthetics*, Vol. 36, No. 3.

Carroll, N. (1998). Art, narrative, and moral understanding. In Jerrold Levinson (ed.), *Aesthetics and ethics: Essays at the intersection.* Cambridge University Press. pp. 126-160.

Carroll, N. (2000). Art and ethical criticism: An overview of recent directions of research. *Ethics*, Vol. 110, No. 2, pp. 350-387.

Carroll, N. (2013). Discussion: Rough heroes: A response to A.W. Eaton. *The Journal of Aesthetics and Art Criticism,* 71: 4.

Dadlez, E.M. (2017). Hume, Halos, and Rough Heroes: Moral and Aesthetic Defects in Works of Fiction. *Philosophy and Literature 41*(1), 91-102.

Eaton, A.W. (2008). Almodóvar's immoralism. In *A. W. Eaton ed., Talk to Her. Routledge. pp. 11-26.*

Eaton, A.W. (2012) Robust immoralism. *The Journal of Aesthetics and Art Criticism*, 70:3.

Eaton, A.W. (2013). Discussion: Reply. *The Journal of Aesthetics and Art Criticism,* 71:4.

Gaut, B. (2007). *Art, emotion, and ethics*. Oxford Scholarship Online.

Hume, D. (1987). Of the standard of taste. In Eugene F. Miller (ed.), *Essays: moral, political, and literary.* Indianapolis: Liberty Fund, p. 246. Cited in Eaton (2012).

Jacobson, D. (1997). In praise of immoral art. *Aesthetics*, Vol. 25, No. 1. University of Arkansas Press., pp. 155-199.

Kaufmann, W. (1974). *Nietzsche: Philosopher, psychologist, antichrist.* Fourth Edition. Princeton University Press.

Kieran, M. (2003). Forbidden knowledge: The challenge of immoralism. In J.L. Bermudez and S. Gardner (ed.), *Art and morality.* Taylor & Francis Group., pp. 56-73.

Larkin, P. (2003). This be the verse. In *Collected poems*. Faber and Faber. London, p. 142. Cited by Stecker (2008).

Nehamas, A. (1985). *Nietzsche: Life as literature.* Harvard University Press. Cambridge.

Nietzsche, F. (1974). The gay science. Trans. Walter Kaufmann. Vintage Books.

Nietzsche, F. (2016). *The birth of tragedy or hellenism and pessimism.* Trans. W.A. Hausmann. Project Gutenberg EBook.

Schpall, S. (2013). The men of *Talk to Her*. Yale Philosophy Department. Online PDF.

Stecker, R. (2008). Immoralism and the anti-theoretical view. *The British Journal of Aesthetics, 48*(2)., pp. 145–161.

Choices and Consequences: A Discussion of Personal Responsibility as a Criterion for Healthcare Allocation

Emma Cox

Abstract

The COVID-19 pandemic has exhausted the resources of many healthcare facilities where the number of patients in need exceeds the number that can receive treatment (Everett, et al. 2021, 932). Clearly, providing treatment to one patient over another carries serious moral implications and therefore should not be done arbitrarily. Pre-pandemic discussions of healthcare allocation have involved social contract theory as a basis for (de)prioritization; under this theory, personal responsibility for one's illness was considered as a relevant criterion. Rawls, in his social contract theory imposes obligations onto individuals who derive benefits from membership in a society (1999, 96). West Virginia's 2006 modified Medicaid program offered enhanced benefits to those who signed a "member agreement" and accepted numerous lifestyle expectations, including submitting to screenings and following health improvement plans (Steinbrook 2006, 753). However, due to the numerous factors, including the social determinants which impact an individual's health, including income, education level, and employment, social contract theories cannot ethically be used to distinguish between patients. As an alternative, utilitarianism has been applied to triage guidelines in the pandemic, supposedly providing a more objective, non-discriminatory basis for treatment allocation which focuses on medical rather than personal factors (Savulescu, et al. 2021, 620). Prima facie, there seems to be a distinction regarding the role of personal responsibility across the two discussed perspectives. Namely, social contract theory directly implies that personal responsibility is a relevant criterion for medical resource allocation, while utilitarianism does not. However, given the inseparability of individuals, their social circumstances, and their subsequent health decisions and outcomes, I contend that both perspectives result in the same moral pitfalls. Further, I argue that personal responsibility ought not to be used as a criterion for healthcare allocation, whether under the application of social contract theory or utilitarianism.

I. Introduction

The COVID-19 pandemic has exhausted the resources of many healthcare facilities and has made necessary difficult decisions for healthcare providers (Everett, et al. 932). Thus, both public health authorities and medical health practitioners must make difficult choices about the allocation of funds and medical resources. While such decisions grant some individuals with potentially life-saving treatment opportunities, they also entail the denial of treatment to others whose livelihood may be contingent on treatment. Given the potential consequences of these decisions, there arises the need to establish principles which can be applied to justify choices of resource allocation. Utilitarian arguments have dominated discussions of healthcare prioritization in the pandemic, with the goal of maximizing the most lives across a large number of patients (Wang 2). However, in cases where there is no discernible difference between patient prognoses or survival chances, this perspective does not provide a basis for patient prioritization.

The idea of individual responsibility for one's own illness has been posed long before the COVID-pandemic, and some healthcare providers assent to using personal responsibility as a decision-making factor for resource allocation. For example, Norwegian and British doctors report that patient de-prioritization for care is warranted for those who engage in smoking, excessive alcohol consumption, and drug abuse (Everett, et al. 936). Further, in a national survey that was conducted in 2006, 53% of Americans reported that they thought it would be "fair" for individuals with unhealthy lifestyles to pay higher insurance premiums, deductibles, or copayments than their healthier counterparts (Steinbrook 753). Similar sentiments are expressed in healthcare promotional campaigns and medical programs involving lifestyle contracts. Thus, there is a clearly established acceptance of personal responsibility for health. While such contractarian perspectives are compelling, I will argue instead, that no patient is more entitled to care than any other, due to the numerous factors which impact an individual's

health, including income, education level, social and community support. Further, I will demonstrate that utilitarian perspectives indirectly assume individual responsibility and involve similar injustices to social contract theory-derived ones.

II. Attributions of Responsibility

Due to accumulating research and epidemiological evidence linking lifestyle factors to heath and disease, health promoters and professionals have adopted the position that behavioral causes are major factors of *preventable* illness (Guttman and Ressler 118). This idea has been incorporated into healthcare promotional campaigns, which establish causal and moral connections between personal behaviors and subsequent health outcomes. By this reasoning, individuals who become ill are those who fail to maintain healthy lifestyles and prevent illness and are thus, morally responsible, and culpable for their conditions. Effectively, the ancient sins of gluttony, sloth, and lust have been replaced by the modern risk factors of overeating, failing to exercise regularly, and engaging in unprotected sex, which hold analogous moral implications for individual agents (Guttman and Ressler 118).

Related to assumptions of personal responsibility for one's own health imposed in public health address, physicians demonstrate agreement that responsibility should be used as a criterion for distinguishing between patients in the face of limited medical resources. In their 2021 study, Everett, et al. examine the sentiments of Norwegian and British doctors on the issue of including personal responsibility for illness in healthcare prioritization decisions. Study participants responded to three vignettes containing descriptions of hypothetical clinical scenarios in which resources are limited and only one patient can be helped. One such scenario posed: "Patient A is a life-long smoker. He grew up on a farm and all his family smoked. He has end-stage emphysema and requires a lung transplant to survive. He is currently smoking… Patient B is a non-smoker but has end-stage emphysema," (Everett, et al. 6). In each hypothetical clinical scenario, most doctors

from both countries answered that they would treat the relatively less responsible patient, or the patient whose lifestyle was not obviously connected to his or her illness (Everett, et al. 9). Thus, in the face of limited resources, physicians consider personal responsibility as a relevant criterion for treatment decisions.

Like the sentiments demonstrated by physicians, the lay public view personal responsibility as a relevant consideration for access to healthcare. Wittenberg, et al. presented survey participants with the hypothetical scenario of a liver transplant decision in which care can be allocated to only one of two patients. While one patient required a transplant due to an inherited factor, the other's liver failure was due to many years of heavy alcohol consumption (Wittenberg, et al. 203). Respondents who believed that those with alcohol-induced liver failure were personally responsible for their disease were more likely to allocate (hypothetical) transplants to the patient with the inherited factor, simultaneously refusing treatment to the alcoholic patient (Wittenberg, et al. 199). Thus, the idea that individuals are morally *culpable* for their illnesses follows the idea that individuals are causally responsible.

The discussed sentiments about personal responsibility for health have been relevant throughout the coronavirus pandemic. Responsibility has been attributed to several identified groups for the virus' proliferation, resulting in sentiments of blame. In the beginning of the pandemic, COVID-19's origin was pointed at the collective actor, 'the Chinese,' who were thought to be responsible for the spread of the virus due to their culinary habits which were characterized as primitive and uncleanly (Barreneche 20). The governor of Veneto, Italy publicly accused, "unlike Italians, the Chinese did not have good standards of hygiene and eat mice alive," (Ivic 424). As the virus was so widely distributed that the Chinese alone could not hold blame, the collective 'posh' were targeted for their vacationing habits which spread the virus across countries (Barreneche 21). Finally, the most widely encompassing group to which COVID- spreading is

attributed is the 'irresponsible' who prioritize their social lives over the well-being of the collective public, by attending social gatherings and refusing to wear masks (Barreneche 21).

III. Social Determinants of Health (SHD)

In recent decades, the public health community's attention has been drawn to social factors as important determinants of individual health outcomes, somewhat diminishing the established role of medical care in shaping health (Braveman and Gottlieb 20). While health outcomes are largely influenced by behaviors, behaviors are strongly shaped by social factors, including income, education, and employment (Braveman and Gottlieb 20). A meta-analysis conducted by Galea, et al. revealed that the number of deaths in 2000 attributable to low education, racial segregation, and low social support were comparable to the number of deaths attributable to myocardial infarction, cerebrovascular disease, and lung cancer (1464). Further, there exists the general trend that health improves incrementally with social position (Braveman and Gottlieb 20). Thus, while there exists a widespread sentiment that individuals who engage in health-related risk behaviors should bear the costs and consequences, imposing responsibility for health onto individuals poses risks for worsening existing social inequalities.

Beyond general health disparities across socioeconomic statuses, there exist racial disparities in COVID-19 outcomes. Through the pandemic, Black, Asian, and minority ethnic groups (BAME) have emerged as more susceptible to higher morbidity and mortality rates than either US or UK white groups (Bentley 1). The CDC found that almost double the amount of Black and Hispanic individuals were hospitalized with COVID-19 than are proportionally represented in the community (Bentley 1). Importantly, social and structural differences predict these disparities rather than racial or genetic differences (Bentley 2). Social and structural inequalities which affect individual vulnerabilities include "exposures through types of employment, whether people are

working in essential transport networks carrying large numbers of people, or in small grocery stores," (Bentley 2). Further, members of BAME communities are at heightened risk for metabolic disorders, including obesity, cardiovascular disease, all conditions linked to higher risk of COVID-19 contraction and poorer outcomes once contracted (Bentley 2).

In addition to disparities in susceptibility to COVID-19 and COVID-19 outcomes, there are disparities regarding vaccine hesitancy (Callaghan, et al. 2). Anti-vaccine advocacy groups, including the Children's Health Defense have targeted African Americans with anti-vaccination messages, potentially contributing to these disparities (Callaghan, et al. 2). Such groups indicate that the COVID-19-vaccine perpetuates the historical pattern of medical abuses against Black Americans in the US, referencing the Tuskegee Syphilis Experiment (Callaghan, et al. 2). These messages promote peripheral trauma and potentially decrease the likelihood that minority groups will pursue vaccination (Callaghan, et al. 2). Affirming this risk, the National Health Interview Survey revealed that in years following the Tuskegee Syphilis Experiment, black men near the Tuskegee area reduced their interactions with outpatient physicians, resulting in a mortality increase (Alsan, et al. 325). In a national survey among Americans, Callaghan, et al. identify the least likely groups to vaccinate were women and Black Americans, with political conservatism also predicting negative intent (5). Importantly for the case of Black Americans, vaccination intentions are reflective of disparities in COVID-19 infection and mortality.

IV. Ethical Theories

While the COVID pandemic is novel and requires some context-specific considerations, there are several ethical perspectives which have been employed to determine the obligations held by physicians towards their patients. Before examining the ethical arguments applied specifically in the pandemic, it is important to understand the

underlying philosophical positions which have been applied across various medical contexts, namely utilitarianism and Rawls' social contract theory.

Given the established weight that individual responsibility holds in discussions of access to healthcare, social contract theory is a relevant perspective in the pandemic context. John Rawls presents the guiding idea for social contract theory as "the principles of justice for the basic structure of society are the object of the original agreement… that free and rational persons concerned to further their own interests would accept in an initial position of equality as defining the fundamental terms of their association," (10). Through this reasoning, Rawls intertwines the concepts of justice and fairness and align both with the interests of individuals and the common good (12). As individuals benefit from being a part of their society, society benefits from having individuals avoid actions which harm the collective good. With these reciprocal benefits come reciprocal obligations; thus, under social contract theory consequences are warranted for those who fail to maintain their obligations to their society.

While Rawls' theory of justice was intended for the general structure of society rather than for a specific contact such as healthcare, some guiding principles of the theory can be analogized to healthcare contexts. For example, Rawls outlines his principle of fairness by defining conditions which give rise to individual obligations (96). He considers an individual to be obligated to comply with a rule of an institution if, first, the institution itself is just. His second condition is that "one has voluntarily accepted the benefits of the arrangement or taken advantages of the opportunities it offers to further one's interests," (96). For Rawls, individuals who derive benefits from a just institution can 'fairly' have their liberties restricted if such restriction yields widespread benefits through the system (96). When analogized to healthcare institutions, the sorts of liberties to be restricted are behaviors which pose health risks, such as smoking tobacco and living sedentary lifestyles. Thus, while Rawls' theory of justice applies to the general structure

of society, the guiding principles which entail individual obligations have been applied in healthcare contexts, to be discussed in the next section.

Another ethical perspective which holds relevance in the pandemic discussion is the consequentialist perspective of utilitarianism. The first notable utilitarian philosopher, Jeremy Bentham, articulates that ethical decisions should be made regarding the amount of pleasure which results, posing also that the number of individuals to whom pleasure or happiness applies must be considered when weighing decisions (Bentham 84). John Stuart Mill, in his *Utilitarianism,* presents the "Greatest Happiness Principle," as the guide for ethical decisions: "actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness," (10).

The general goal of utilitarianism as maximizing benefit for the greatest number of people provides some ambiguities which hold with regard to medical equipment and treatment during the pandemic. For instance, Mill furthers Bentham's value for quantity of pleasure by providing that quality of pleasure must matter as well (Mill 11). He reasons that higher quality of pleasure can be found in only intelligent beings, whose experiences surpass those which can be attained by lower animals; thus, for Mill, the pleasures of intellectual discovery rank over the pleasures provided by eating something delicious (12). He explains that "few human creatures would consent to be changed into any of the lower animals," and that, even amidst the heightened risks of suffering felt by rational beings, humans "can never really wish to sink into what he feels to be a lower grade of existence," (2). This regard for quality of pleasure raises important considerations for utilitarian arguments in the pandemic, potentially presenting disadvantages for patients of low cognitive capacities related to disability or physical condition.

Both utilitarian and Rawls' social contract theorist perspectives provide direction for navigating healthcare allocation decisions; however, both also entail issues

of inequity and inequality which deserve careful comparison and consideration. For instance, Mill's high regard for quality of pleasure may result in a de-prioritization of individuals with cognitive disabilities. Similarly, even if quality of life is dismissed, Mill's utilitarianism may result in other forms of discrimination. For instance, vulnerable groups may stand to benefit less from treatment than their healthier counterparts, given that chronic illnesses reduce life expectancy and therefore reduces the relative utility of a treatment (Savulescu, et al. 623). Similarly, social contract theorist perspectives imply ethical problems of blaming those who suffer from social inequalities which negative health outcomes (Steinbrook 755). While self-interested individuals should supposedly avoid risk-decisions which may harm their society and their own resources, those who reside in lower social positions do not enjoy the educational, structural, and monetary benefits which facilitate healthy behaviors.

V. Applied Theories in Healthcare Contexts

Prima facie, there seems to be a distinction of the role of personal responsibility across the two discussed perspectives, namely, Rawls' social contract theory directly implies that personal responsibility is a relevant criterion for medical resource allocation, while utilitarianism does not. Utilitarian perspectives appear to be 'fairer' in that they do not consider causes of illness or invoke blame to individuals. However, both perspectives result in the same moral pitfalls considering the inseparability of individuals, their social circumstances, and their health decisions and outcomes. Thus, the argument against employing personal responsibility as a criterion for medical resource allocation extends, not only to social contract perspectives, but also to utilitarian ones.

VI. Social Contract Theory in Healthcare

One example of how social contract theory may be applied in healthcare appears in the 2006 re-design of West Virginia's Medicaid program to incorporate personal responsibility as a qualifying factor for access to healthcare (Steinbrook 753). Under the

new plan, most low-income healthy adults and children received reduced basic benefits. However, by signing and adhering to the "Medicaid Member Agreement", enhanced benefits could be obtained. These benefits include all mandatory services in addition to wellness-focused age-appropriate services, such as diabetes care, cardiac rehabilitation, tobacco-cessation programs, education in nutrition, chemical dependency, and mental health services (Steinbrook 754). There is a clear differentiation between the basic and enhanced plan and thus, a clear incentive to accept personal responsibility for health. For instance, while the basic plan only allots four prescription refills per month, the enhanced plan provides no limitations. To keep enhanced benefits, members must successfully comply with four responsibilities, including keeping medical appointments, receiving screenings, taking prescribed medications, and following health improvement plans (Steinbrook 754).

The Commissioner of the Bureau for Medical Services in the West Virginia Department of Health and Human Resources articulates the main goals of the program's redesign, as "provid[ing] members with the opportunity and incentive to maintain and improve their health," (Steinbrook 754). However, there are legitimate reasons for which members may not comply with enhanced plan conditions, including poor physician-patient communication, side effects of medication, impractical advice regarding job responsibilities, transportation, childcare, psychiatric illness, cost, complex recommendations, and language barriers prohibiting understanding of recommendations (Steinbrook 755). Further, the patients in most need of enhanced services, such as diabetes care, education in nutrition, and chemical-dependency and mental health services, may be those with the most difficulty complying.

Given that health related behaviors are significantly linked to social factors, including education, employment, and income (Braveman and Gottlieb 20), imposing responsibility for health onto individuals rather than social inequalities would not likely

improve health outcomes or modify health decisions. Thus, those with the most to gain from the enhanced services plan are likely to be those who are excluded from it. Despite the benevolent intentions of the plan, there is a risk for rewarding those with fewer needs for enhanced benefits and marginalizing those who are most vulnerable

Some less explicitly social contract theorist positions have been incorporated in healthcare discussions, though they result in the same risks to equity and equality as does the WV Medicaid program re-design. For instance, Alena Buyx argues that personal responsibility can ethically be used as a criterion for rationing decisions, proposing liberal egalitarianism to reconcile the negative associations with responsibility-based resource allocation, such as libertarian perspectives. For instance, libertarian healthcare proponents argue that individuals have the right to decide on how to spend their funds according to their life plans and reject any mandatory redistribution of personal funds to social programs (Buyx 871). However, as Buyx points out, under such a healthcare system, large portions of the population would be left without public support in cases of illness (872). Conversely, proponents of communitarian theories of justice argue that the common good outweighs the importance of individual preferences (Buyx 871). Thus, preventative and rehabilitative treatment for the public should replace expensive treatments for the few in the pursuit of a healthier population. However, individuals who become ill despite preventative and rehabilitative efforts would be considered burdensome to the common good due to their need for expensive treatment (Buyx 872).

As an alternative, Buyx proposes liberal egalitarianism which balances the needs and preferences of individuals with the need to support societal institutions to the end of protecting equality of opportunity (Buyx 872). This perspective encompasses the principle of solidarity, a sense of togetherness between the members of a society. Togetherness, in this context, entails being part of a system deemed precious and important and therefore, requiring members to support it and actively attempt to avoid

harming the system (Buyx 872). Thus, a liberal egalitarianist medical system would require its members to act responsibly regarding their health. However, to avoid the discussed consequences of libertarian and communitarian healthcare systems, Buyx proposes that personal responsibility only serve as one criterion among many in a matrix used for care allocation (873). Additionally, she maintains that even in cases of personal responsibility for illness, baseline healthcare provisions are necessary (872). Finally, she proposes that incentives should be offered for those who engage in programs designed to combat problematic health behaviors such as smoking, sedentary lifestyles, or bad diets.

Finally, Buyx acknowledges that if personal responsibility were to be employed as a criterion for healthcare access, efforts would have to be made to change the "toxic environment" and diminish social impact on health behavior (873). For a person to retain responsibility for herself, she must possess adequate knowledge and health literacy to make informed decisions. Thus, improving widespread education about health maintenance are necessary before personal responsibility can ethically be employed to make treatment allocation decisions. Further, Buyx acknowledges the problem of social stratification of health behaviors, which could be worsened if personal responsibility were to be incorporated into healthcare access decisions, by imposing burdens onto already vulnerable groups (874). Despite the problems attached to imposing personal responsibility, Buyx's final resolve is that personal responsibility will likely improve health and therefore ought to be placed as a consideration in healthcare access.

While Buyx paints a hopeful image of a healthier society, current social conditions and health disparities prevent any ethical implementation of such a program. For instance, Andreas Albersten presents the criticism to liberal egalitarianism that it is "not sufficiently attentive to the complex relationships between social circumstance and health outcomes," (564). Albersten demonstrates that the metaphysical debates about causation and responsibility are inevitable components of the healthcare discussion, as

many behaviors are contingent on social determinants in health, including where people live, whether they are employed, and their general socio-economic positions. Thus, imposing personal responsibility cannot be equitable or fair due to the stratifications in social conditions which impact behavior and subsequent health outcomes.

VII. Utilitarianism in Healthcare

During the initial months of the pandemic, the threat of medical resource exhaustion grew. As the number of patients in critical condition exceeded the number of ventilators and ICU beds available, healthcare providers were forced to choose to treat some patients and not others. However, the US Department of Health and Human Services promised that "persons with disabilities, limited English-speaking skills, or needing religious accommodations should not be put at the end of the line for health services during emergencies. Our civil rights laws protect the equal dignity of every human life from ruthless utilitarianism," (Savulescu, et al. 620). Utilitarianism as a moral theory is often criticized as a ruthless theory which reduces individuals to their utility and therefore uses them as means to certain ends (Savulescu, et el. 621). However, despite some of the associations with the ethical theory, the scope of the pandemic necessarily places many lives at stake and presents difficulties in justifying focusing on individual-rather than population-level benefit.

In their comparative analysis of the national and international triage policies designed for the pandemic, Susanne Jobges, et al. determine utilitarianism to be the prominent ethical perspective worldwide (949). However, the goal of maximizing benefit does not afford clear criteria which can be employed to distinguish between patient prospects. For instance, maximizing benefit could entail maximizing the number of lives, regardless of prognosis, comorbidities, or age (Jobges, et al. 949). Conversely, it could entail maximizing the number of life years saved, which would privilege those with stronger survival prospects and greater life expectancies. Further, maximizing benefit

could mean focusing on quality-adjusted life years, which favor those with a capacity to live long, independent lives. This may necessarily incorporate some forms of discrimination towards those with cognitive or physical impairments, as impairments could limit the kinds of benefits that can be enjoyed after treatment (Jobges, et al. 957). Maximizing benefit also necessitates considerations for those of "instrumental value," such as healthcare workers who endanger their own lives while potentially saving many others. While some of these distinct kinds of benefit maximization may be combined, some choices are necessarily mutually exclusive. For instance, comparing a young patient with a severe cognitive impairment but otherwise good health and a much older patient with no cognitive impairments, either quality-adjusted life years saved, or mere quantity of life years saved must be chosen as a basis for prioritization.

Whichever conception of benefit maximization is accepted, there are necessary ethical implications which follow. For instance, if maximizing benefit is interpreted to mean maximizing the number of lives, regardless of prognosis, comorbidities, or age, the result could be a massive preventable loss of life (Savulescu, et al. 620). Employing such a blind method of treatment prioritization would likely entail that individuals with low survival chances are treated in favor of those with many life years to gain, a consequence which would be difficult to justify under the mere premise of equal and equitable access to treatment. Using this blind method would likely result in the loss of lives which could have been prevented if patient health conditions and survival chances were considered. However, if the decision-making aim becomes maximizing the number of life years saved, there necessarily arise issues of inequity and inequality associated with health disparities across socioeconomic conditions.

Given the nature of decisions which must be made during crises such as the pandemic, Salvulescu, et al. first consider some of the utilitarian 'rules-of-thumb' employed (623). The dominating rule for utilitarians is number; thus, when allocating

medical resources, the aim should be to maximize the number of lives saved (Savulescu, et al. 623). Salvulescu, et al. propose several triage scenarios and derive several sub-rules-of-thumb which support maximizing benefit across numbers. First, they consider a patient with a 90% survival chance with another who has only a 10% survival chance. In this case, the clear intuitive utilitarian position favors the patient with a higher likelihood of survival, given that treating the riskier patient may result in two lives lost (623). Savulescu, et al. also consider the importance of resources in weighing such triage decisions. For instance, if one patient will likely require ventilator treatment for four weeks, while the other would likely benefit after only one week, there is a utilitarian basis for treating the latter patient and making the ventilator available for others in need, since this will result in more lives saved (623)

Another important criterion which comes into play in triage decisions aimed at utilitarian outcomes is life expectancy (Savulescu, et al. 623). The end of maximizing benefits is impacted more by individuals whose lives are saved by longer rather than shorter periods of time. Thus, utilitarian principles tend to favor the young in triage decisions; though if a younger person held a lower life expectancy due to some non-age-related factor, the opposite decision would be justified. While age, in many cases, is tied to life expectancy, Savulescu, et al. maintain that this criterion is not an explicit form of ageism because the length of the benefit is the justification for such choices.

Beyond simply quantities of lives and life years, utilitarians also consider quality of life. While this poses concerns for protections of vulnerable groups, such as those with cognitive or physical disabilities, the goal of benefit maximization necessarily entails regard for life quality. To exemplify this reasoning, Savulescu, et al. propose a treatment decision between a patient who works full time and possesses all his mental faculties and a patient whose end stage dementia predicts that she will be rendered unconscious soon (623). While both patients would likely survive the treatment and

probably stay *alive* for comparable amounts of time, it would be difficult to make the case that both patients would derive the same benefit from the treatment. Further, it would probably be equally difficult to argue that the precious medical resources would be best spent on the cognitively impaired patient.

Savulescu, et al. further point out that utilitarianism is often in direct conflict with the principle of responsibility in healthcare decisions (625). This is because for utilitarians, intentions do not matter. Utilitarians reject "all direct consideration of causal contribution to illness, and indeed, any backward-looking considerations," (625). Thus, though personal responsibility may pose concern for an individual whose lifestyle of overeating caused diabetes, for utilitarians, it is only relevant as it impacts survival likelihood and life expectancy. While using medical criteria in resource allocation decisions may satisfy utilitarian goals of maximizing quality life years saved, doing so necessarily implies personal responsibility for health.

VIII. Conclusion

Considering the two applied ethical perspectives aimed at justice and fairness in healthcare—namely, social contract theory and utilitarianism—, there arise disquieting implications regarding social inequalities. Healthcare conceptions of Rawls' social contract theory directly attribute personal responsibility for health to individuals, making healthcare availability reflective of the risks associated with their lifestyle factors. While this sort of system seems to empower individuals with the ability to determine their healthcare options, empirical evidence suggests that behavioral factors are highly associated with socio-economic factors. Thus, social contract theory- derived healthcare systems pose the ethical risk of blaming individuals for the social inequalities they are suffering from.

Utilitarian perspectives focus on medical criteria rather than personal lifestyle considerations, thus providing a more objective way of allocating healthcare. Under these

guidelines, patients are evaluated in terms of the benefits they may derive from medical treatment compared to other patients in need. While this seems to eliminate the victim-blaming problem of social contract systems, utilitarian guidelines result in the same disfavoring of the already-vulnerable. The same end is met whether a patient is denied access to a ventilator because he smoked cigarettes for fifty years or because his lung disease worsens his life expectancy compared to other patients. The implications of socio-economic factors on individual lifestyles are similar to their implications on health factors, such as metabolic disorders, obesity, and cardiovascular diseases, all comorbidities associated with negative outcomes with COVID-19.

The reviewed social scientific literature presents a bleak, deterministic model which related individuals to their social circumstances, lifestyles, and health outcomes. Whether through Rawlsian social contract theory or utilitarianism, the 'fair' and the 'just' allocation of resources only pose benefits for the privileged. Thus, there arises the need to allocate resources in ways that favor the most vulnerable members of social systems. Given the drastic social inequalities which persist through COVID-19 outcomes, ethicists have proposed that those who are 'most unfairly exposed to SARS2, such as poorly paid worders in nursing homes… [or] prisoners or undocumented workers held in crowded detention centers," (Pence 83). This sort of resource allocation would work against social inequities and inequalities, potentially diminishing the health disparities that exist across racial, ethnic, and economic lines.

Another way that healthcare allocation could be used to work against inequalities is shown in a New York policy that allows nonwhite race or Hispanic ethnicity to be a consideration when dispensing anti-viral treatments which are limited in supply (Woodward and Klepper 1). This policy is aimed at steering treatments to those who at the most risk of severe disease from coronavirus, citing that "long-standing health and social inequalities make people of color more likely to get severely ill or die from the

virus," (Woodward and Klepper 1). Such policies could also be implemented with respect to disparities outside of COVID-19 outcomes, providing priority for surgery or organ donation for those whose socio-economic factors place them at risk for negative health outcomes.

Though both utilitarianism and Rawls' social contract theory both attempt to provide justice and fairness in the face of limited medical resources, both fall short due to existing health disparities and social inequalities. While the principles of each theory may be ethically acceptable in a world of widespread social equality, neither can be used while such injustices persist. Thus, healthcare allocation must be aimed at helping the most vulnerable groups in society until their social circumstances no longer pose such bleak implications for their health.

Works Cited

Albertsen, Andreas. "Taxing Unhealthy Choices: The Complex Idea of Liberal Egalitarianism in Health." *Health Policy* 120.5 (2016): 561-566.

Barreneche, Sebastian M. "Somebody to Blame: On the Construction of the Other In the Context of the Covid-19 Outbreak." Society Register 4.2 (2020): 19-32.

Bentham, Jeremy, and John Stuart Mill. *Utilitarianism and Other Essays*. Penguin UK, 2004.

Bentley, Gillian R. "Don't Blame the BAME: Ethnic and Structural Inequalities in Susceptibilities to COVID−19." *American Journal of Human Biology* 32.5 (2020)

Buyx, Alena M. "Personal Responsibility for Health as a Rationing Criterion: Why We Don't Like it and Why Maybe We Should." *Journal of Medical Ethics* 34.12 (2008): 871-874.

Callaghan, T., Moghtaderi, A., Lueck, J., Hotez, P., Strynch, U., Dor, A., Fowler, E., Motta., M. "Correlates and Disparities of Intention to Vaccinate Against COVID-19." *Social Science & Medicine (1982)* (2021).

Everett, Jim A. C., Maslen, H., Nussberger, A., Bringedal, B., Wilkinson, D., Salvuescu, Julian. "An Empirical Bioethical Examination of Norwegian and British Doctor's Views of Responsibility and De-prioritization in Healthcare." *Journal of Medical Ethics* (2020).

Galea, S., Tracy, M., Hoggatt, K. J., DiMaggio, C., Karpati, A. "Estimated Deaths Attributable to Social Factors in the United States." *American Journal of Public Health* 101.8 (2011): 1456-1465.

Guttman, N., & Ressler, W. H. (2001). On Being Responsible: Ethical Issues in Appeals to Personal Responsibility in Health Campaigns. *Journal of Health Communication*, *6*(2), 117-136.

Ivic, Sanja, and R. Petrović. "The Rhetoric of Othering in a Time of Pandemic: Labeling as a Foreign Virus in Public Discourse," Kultura Polisa (2020): 421-433.

Jöbges, S., Vinay, R., Luyckx, Biller-Andorno, N. "Recommendations on COVID−19 Triage: International Comparison and Ethical Analysis." *Bioethics* 34.9 (2020): 948-959.

Mill, John Stuart. "Utilitarianism (1861)." *Utilitarianism, Liberty, Representative Government* (1979): 7-9.

Pence, Gregory E. Pandemic Bioethics. Broadview Press, 2021.

Rawls, John. *A Theory of Justice: Revised Edition*. Harvard University Press, 1999.

Savulescu, Julian, Ingmar Persson, and Dominic Wilkinson. "Utilitarianism and the Pandemic." *Bioethics* 34.6 (2020): 620-632.

Steinbrook, Robert. "Imposing Personal Responsibility for Health." *New England Journal of Medicine* 355.8 (2006): 753.

Wang, X. "The Fairness of Ventilator Allocation During the COVID-19 Pandemic." *Bioethics* (2021) 1– 9. https://doi.org/10.1111/bioe.12955

Woodward, Calvin, and David Klepper. "Ap Fact Check: Trump Seeds Race Animus with COVID ..." *U. S. News*, 16 Jan. 2022, https://www.usnews.com/news/politics/articles/2022-01-16/ap-fact-check-trump-seeds-race-animus-with-covid-falsehood.

Emma graduated from the University of Southern Mississippi in Spring of 2022. She double-majored in Philosophy and Communication Studies and completed her *Sapere Aude* submission in fulfillment of her Philosophy Capstone requirements. Additionally, she presented the paper at Mississippi Academy of Sciences in the History and Philosophy of Science division, winning the undergraduate presenter award. Now that she has graduated from her undergraduate institution, she will attend Clemson University for the MA in Communication Studies, where she will focus on research in health communication and message design. While she is no longer formally studying philosophy, she remains committed to pursuing ethical questions in applied contexts, particularly including healthcare settings.

Sadism in the Bedroom: Metaethical Reasons to Prefer Kantianism to Utilitarianism

Andy J. Baldassarre

Abstract

Through the consideration of sadistic sex acts between consenting parties, a case can be constructed which shows the inability of Utilitarianism to accommodate some acts even when all affected parties are consenting and acting rationally. This may be cause to favor Kantianism as a moral theory.

I. Introduction

Identifying *the* True Moral Theory is in many ways the quintessential goal of moral philosophy. To this end there are many metaethical tools available for trying to identify what moral principles may be "true" from the perspective of the universe (if any). One such tool is the use of moral intuition to evaluate the implications of applying a given moral theory. If moral intuition and the casuistry of applying an ethical principal are in conflict, then there is cause to scrutinize both the principal and the intuition. It stands to reason that the more robust of the two is likely to be closer to the truth. These dilemmas can be operationalized through the careful construction of case studies which highlight the incongruities therein.

Something like a Kantian or neo-Kantian view of morality might more accurately describe what is moral from the perspective of the universe than a Utilitarian moral theory. To demonstrate this, we can first apply Kant's usage of rational thought to identify that which is right and wrong to further clarify personal duties that exist for individuals, insofar as they are a member of a larger subset of all moral agents. By exploring the personal duties of lovers, a case can be constructed in which, assuming a condition of consent is satisfied by all parties, Utilitarianism is unable to permit actions which are both permissible under Kantianism and agreeable to all affected parties. Such a case highlights the limitations of acting to maximize utility in the face of contraindications from potential moral duties. The case described examines an explicit exchange of pain for pleasure with a net loss of utility as may result from a sexual encounter between a sadist and a non-masochist.

II. Kant and Personal Duties

Kant asserts that morality proceeds from acting in accordance with our duties. Duty, as Kant describes, is what we ought to do such that our behavior is

rationally consistent[98]. A meaningful conception of duty, however, could be extended in accordance with an agent's identity. In other words, a rationally consistent world may entail the establishment of a distinction between the duties of agents in accordance with what might reasonably be expected of them. Acting according with *universal* duty compels us to act in accordance with *personal* duty. The duties of a parent, a physician, an employee, as examples. We inherit additional duties in our various roles dictated by our relationships to others, and when we consent to taking on a new role, we necessarily consent to taking on the duties contained therein.

Personal duties are context dependent – they are those things for which it would be rationally consistent for all agents with a shared identity to similarly do. The duties of a parent are those which all parents ought to do. The duties of a physician are those which all physicians ought to do[99].

Another key aspect of Kant's ethical teaching is that an agent must never treat another person as a mere means. The word "mere" is doing a great deal of work in this phrase, and a tremendous amount of proverbial ink has been spilled debating the precise meaning of this word. Operationally for our purposes here, the distinction between "means" and "mere means" can be summarized as follows: an agent, A_s, is being treated as a means in any such transaction as they are acting in service to another agent, A_a, but they are not merely a means if this action is undertaken by A_s willfully in accordance with their personal duties or A_s has consented to engage in this transaction with A_a because it fulfills some end of A_s's.

---

[98] Rational consistency is a condition that may apply to an action or belief. Rational consistency is judged on the basis of several conditions. The most relevant for our purposes are Kant's maxim of Universalizability. If an action continues to be meaningful if everyone did it all the time then it is a rationally consistent act. (Kant)

[99] We might refer to this as the "Any True Scotsman" test. If all the elements of a set share a property, then every individual element of the set must necessarily possess that property. Consider the personal duties of a judge as an example. Any duty which is the duty of all judges, such as being apprised of the law, must therefore be the duty of every judge. That which is a duty of any true Scotsman is the duty of every true Scotsman.

A child that eats the food provided by their parents without "contributing" is not using their parents as a mere means because it is the personal duty of a parent to support their child, and these parents have willfully done so. Alternatively, the barista at one's local coffee shop is not a being used as a mere means because while the customer uses the barista as a means to acquire their order, the barista uses the customer as a means to remain employed (presumably among the barista's goals).

III. What Kant Can Teach Us About Being Ethical Lovers

Let us now consider the personal duties that accompany the role of lover[100]. Specifically, in the context of a sexual relationship, a lover has a duty to their partner or partners. I propose the personal duty of a lover is as follows:

PDL1: A lover has a Kantian duty to satisfy their partner(s).

We can verify this with a rational test under the condition of universalizability. It is rationally consistent to suppose that all lovers ought to sexually gratify their partners. Ought implies can, of course, so the burden might be reduced to the statement that "all lovers ought to *try* to sexually gratify their partners" instead. In this form the duty of a lover can be reformulated as follows:

PDL2: A lover has a Kantian duty to try to satisfy their partner(s) in a sexual capacity.[101]

This is to say their actions should be motivated in accordance with their sense of duty. Consent becomes an essential component of appraising the morality of actions in this context. While it is perhaps a duty of lovers to try to be the best lover that

---

[100] "Lover" is a loaded term. Here I am using the title to denote an appropriate and enthusiastic sexual partner to someone in which the relationship between them (or some facet of it) centralizes sexual desires. Being a lover neither entails nor precludes having other relationships with the same individual. Two agents can be spouses without being lovers, two agents can be both lovers and friends, two agents can be strictly lovers. In this paper it will be used to signify two agents who are routinely engaged in sexual relations, independent of other roles they may play in each other's lives.

[101] This is not to say that all people ought to be having sex all the time. This is a much weaker claim. All this says is that when one agent enters into a relationship with another such that they are now that agent's lover they assume additional responsibilities which follow from the personal duty, PDL2.

they can be with their partner, should the partner push them to do something to which they do not consent[102] the partner is now using their lover as a mere means – merely a means for their own sexual gratification. The Kantian understands that the partner has done wrong, they have acted unethically.

Proceeding only from a Kantian framework it can be reasoned that ethical behavior in sexual relationships must feature at least two elements: one being consent and the other being an earnest attempt on the part of lovers to satisfy the desires of their partners. A dutiful lover (and crucially, an ethical lover) does not simply phone it in.

IV. The Sadist and The Good Sport

Consider the following case. Two parties have entered into a relationship. Sam (the sadist) and Winnie (the willing one) are in a very happy romantic relationship. So happy, in fact, that neither one of them sees sex as registering even a slight factor on the quality of their relationship. They are going to continue to be together faithfully and happily, come what may. That being said, they both enjoy sex with each other and derive pleasure from sex acts. Sam and Winnie do not, however, have the same sexual preferences. Sam derives pleasure from sadistic acts. If Sam inflicts pain on Winnie in the bedroom Sam will increase their own personal satisfaction. Winnie is not a masochist and derives no pleasure from having pain inflicted upon them. Winnie is, however, a happy participant in Sam's desires because Winnie seeks to be the best lover they can be. Winnie needs no coercion, no duress, and feels completely at liberty to refuse Sam's requests (perhaps even does refuse them periodically).

[102] Consent cannot be so weakly defined as affirming a willingness to do something. Consent is a robust and multifaceted condition. An agent under duress or similarly pressured to act a certain way is not satisfying an authentic condition of consent. There also exist epistemic conditions to consent – an agent cannot consent to an act about which they have limited or false information. Hereon the word "consent" indicates a broad and authentic condition of consent in which all parties are informed and earnest in their willingness.

A Utilitarian sees a very simple mathematical problem here. Sam will gain pleasure by inflicting pain (P_S), Winnie gains no pleasure, nobody loses their autonomy so it cannot be said that some abstract harm associated with the violation of consent enters into the calculus, and Winnie experiences physical pain (N_W). Assuming that Sam proposes a sadistic sex act, Winnie can either approve or disapprove (simplified to a binary set of responses for the purposes of this argument). So long as $|P_S| - |N_W| > 0$ the addition of this sadistic action will cause a net increase in the pleasure derived from sex between Sam and Winnie. Should $|P_S| - |N_W| < 0$ then there is a net decrease in the pleasure derived from such a sex act. Therefore, the Utilitarian says, the action is easily morally evaluable – so long as Sam gains more pleasure than Winnie experiences pain this is a good action. If Winnie experiences more pain than Sam experiences pleasure, then it is a bad action.

This last statement proves troubling. Winnie may consent to a sex act that they fully know will cause marginally more pain than Sam will gain pleasure (perhaps this knowledge comes from past experiences). Why then would Winnie agree to such an act? Well, Winnie might reason, this does hurts them, but not terribly so and not such that they have ever felt or expect they ever will feel unsafe. Moreover, they think it is right to do all that one *reasonably* can as a lover to please their partner. Winnie is a Kantian and believes PDL2 is indeed true. The Utilitarian is now left claiming an act to which all impacted parties are assuredly consenting is immoral.

<u>The Utilitarian has several options on how they might proceed:</u>

1. *No clarification is needed! This action is immoral, what's the issue there?*
2. *Consent is necessary, sure, but it is not a sufficient condition for moral permissibility.*

3. *This case is fundamentally flawed! A rational agent would surely not consent to their own harm unless they benefit in some other way. Details are being left out somewhere.*

Each of these responses fail to account for general moral intuitions. Allow me to work through them in order.

Perhaps the problem with the first response from a Utilitarian is clear enough on its face. That response is truthfully the Orthodox Utilitarian view in which dispassionate calculations of pain and pleasure are not just guides for comparing choices, but rather that calculus is strictly and singularly an arbiter of moral rightness. Consent is irrelevant to questions of right and wrong. Under this view operations like genocide and slavery are not just permissible, they may be obligatory in certain cases[103]. These conclusions seem repugnant on their faces[104].

A more charitable interpretation might look like some kind of satisficing form of Utilitarianism. The Utilitarian putting forward the second response concedes that the violation of consent is wrong (maybe even strictly wrong), but that's because there are greater pains associated with one's consent being violated. Physical pain, the Utilitarian may argue, is paltry compared with the suffering of losing autonomy. That being said, just because affected parties are willing to engage in an action does not mean that action is right. Consent – for the proponent of the second argument – does not make otherwise wrong things right.

This view is substantially more robust. Surely some things are wrong *prima facie* and no amount of enthusiasm on behalf of those involved can make things otherwise. There exists a pragmatic problem with this conception of rightness and

---

[103] Such a case would have to produce sufficient utility for those perpetrating atrocities to offset the disutility associated with those who are suffering. Aristotle's description of the leisurely life serves as an example. (Aristotle)

[104] It is beyond the purview of this paper to demonstrate that genocide and slavery are strictly immoral. That being said, if the claim that a moral theory permitting – and even requiring – slavery and/or genocide might be flawed is a contentious one, then it may be beyond my ability to convince you of anything in a succinct form.

wrongness. One of the major appeals of Utilitarianism is the consideration of the suffering of victims of bad actions. Too much suffering because of a largely good act is reason for pause on behalf of the Utilitarian. Bad situations are often not made easier for victims of good intentions gone awry, and the Utilitarian completely accounts for this suffering. In many ways the Utilitarian view seems compassionate in such cases, it concerns itself with the experiences of victims.

The problem for Utilitarian thinking in the case of Sam and Winnie is that there are no victims. Sam is certainly benefitting from the inclusion of sadistic acts, and Winnie is by no means a victim. What can be said of Winnie is that they are experiencing pain. Winnie being in pain, however, does not make Winnie a *victim*. This is just as a surgical patient is not a victim of the surgeon's - even when the surgery is unsuccessful and the Utility is strictly negative the word "victim" is not applied. If the Utilitarian wants to argue that consent is not a sufficient right making feature about an act that they have deemed wrong, then the Utilitarian must first demonstrate that there is a victim of said wrongdoing. If they cannot identify a victim then "wrongness" in such a case is purely abstract, verging on inconsequential (a conundrum for the Utilitarian) – and moreover Sam and Winnie would probably find it distressing to be told that their consensual sexual practices are "wrong".

The psychology of someone like Winnie seems inscrutable. The Utilitarian proposing the third response is fixating on the irrationality of an agent consenting to their own harm if it really is a net negative. Winnie must derive some other benefit not previously described.

Utilitarianism will still come up short in this case, but let us now modify the equation $|P_S| - |N_W| = \sum_U$ and add additional terms accounting for the benefits to Winnie so as to come up with $|P_S| + |P_W| - |N_W| = \sum_U$. Here the new term, $|P_W|$, indicates the utility value of Winnie's benefits. The Utilitarian could rely on benefits that are

psychological. These benefits could be the empathetic joy received from causing a loved one pleasure, or perhaps the sense of security gained from doing an act that contributes to the maintenance of a relationship that one finds beneficial.

Winnie's consent (a necessary condition for their not being a "victim") must indicate that $|P_W| - |N_W| \geq 0$ and therefore $\sum_U \geq |P_S|$. Any other explanation must mean Winnie is willing to consent to a state in which they are worse off. That would appear to be irrational.

Firstly, constructing the case mathematically is trivially easy. It is a given that sex has no bearing on the security of Sam and Winnie's relationship. Winnie is therefore only deriving additional pleasure insofar as they can find pleasure in satisfying Sam. Even then, this is not the only way to satisfy Sam. If Winnie can pleasure Sam (albeit to a lesser degree) in a manner such that they are also sexually gratified without experiencing any pain and find pleasure in that (albeit to a similarly reduced degree) the net utility would still be higher. We can conclude then that $|P_W|$ is relatively trivial compared to $|P_S|$ and the case can be reformulated endlessly to increase the value of $|N_W|$ (one more stroke of a whip, one more poke of a pin, etc., until $|P_S| + |P_W| - |N_W| < 0$).

If such a case is constructed, we may be inclined to regard Winnie as an irrational agent. Why would any agent knowingly enter into an agreement in which they are decidedly worse off with no hope for future gain? I propose that they are acting in accordance with their duty.

Suppose a stranger hands a dollar to an unhoused individual asking for spare change. That stranger certainly has no expectation that she will be later rewarded for this. She likely acted on instinct, doing what she believes to be the right thing to do, and will not dwell on this interaction as she goes about the rest of her day. Her behavior

need not rely on some kind of internal psychology that rewards her. A sense of gratification is not a requirement for explaining her behavior as completely rational[105].

Her willingness to do right acts, even at strictly personal cost, should not be a sign to us she acts irrationally, rather that she has a well developed sense of moral duty. Not all actions in the world need to be transactional, sometimes we may simply do right things because they must be done. This is true for Winnie as well. Their willingness to engage in sex acts that are strictly for the benefit of another is not a sign that Winnie is irrational. They may consider it a part of their duty, as Sam's lover, to facilitate Sam's sexual gratification however best they can. Here our case leaves us with a set of actions that rational agents could consent to from which there are no victims, even though the action produces less utility than inaction. Utilitarianism, as it is most traditionally explained, would appear deficient in its ability to make sense of this dilemma. Kantianism, from which an understanding of personal duty and the inherent value of consent arises, makes such victimless actions such as these clearly defensible.

V. Conclusions on Kant, Consent, and Kink

A single hypothetical case does not disqualify an entire moral theory. What is evident, however, is that Utilitarianism does not seem a sufficient moral theory in a reasonably plausible case as the one above. Utilitarianism, even in a satisficing form with a stipulation that consent is necessary or otherwise intrinsically valuable, prohibits or discourages certain sex acts even if all parties are happily participating while fully consenting and fully able to consent. This is true even when there are no identifiable victims who would consider themselves harmed. It seems counterintuitive to claim that when all agents – acting rationally in accordance with their own duties – affected by a

[105] Kant observes that while it is perhaps ideal for someone to enjoy and feel good about acting rightly, an agent's feelings on the matter are ultimately immaterial. The moral law is strict and acting rightly is not governed by feelings on the matter. Charitable giving, for example, continues to be right (and rational) whether or not the giving agent feels positively about their charitable act.

given choice or set of choices fully consent to those choices and have no regrets after seeing the outcome of those choices they can still have somehow acted wrongly.

Kant's moral theory, that right acts are those taken in accordance with our duties, neatly accounts for the general intuition that consent is valuable, and coercion is wrong without requiring any qualifications. It can also explain why someone might act righteously even at personal cost without needing to claim some deficiency of reason or rationality on behalf of the agent. While this by no means demonstrates Utilitarianism is categorically false, the fact that cases in which apparently nothing immoral is occurring can be flatly impermissible by Utilitarianism would indicate some fundamental flaw(s) in the theory – or at least a major blind spot.

Works Cited

Aristotle. Aristotle's Politics. Oxford :Clarendon Press, 1905.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. 1785.

Kneller, Jane. "Kant on sex and marriage right." Guyer, Paul. *The Cambridge Companion to Kant and Modern Philosophy*. New York: Cambridge University Press, 2006.

Varden, Helga. "Kant on Sex. Reconsidered. -- A Kantian account of sexuality: sexual love, sexual identity, and sexual orientation." *Feminist Philosophy Quarterly* (2018).

Andy Baldassarre is an American philosopher with particular interests in clinical and applied ethics, moral realism, and metaphysics. He is an incoming Master's student at the University of Houston where he will be a Teaching Assistant for the Philosophy Department. He spends much of his free time working with honeybees, beekeeping and studying the needs of bees and other pollinators. He received his Bachelor's Degree in Mechanical Engineering in 2022 from Northeastern University where he double majored in Philosophy.

The Private Language Argument and Mind-Body Dualism: A Reassessment

Álvaro R. G. Barredo

I. Introduction

The Private Language Argument (PLA) turned Wittgenstein's *Philosophical Investigations* into a source of heated debate due to the overarching implications it has for the way we have traditionally understood the endeavour of philosophy, casting doubt on our ability of introspection, and even threatening to render such a concept unintelligible. In this essay, I shall explore the bearing the PLA may have on the philosophy of mind, and, more precisely, whether it is a conclusive objection to mind-body dualism – a concern that has been raised before in the literature (Villanueva, 2: 30). The PLA, under a common interpretation, reduces all putative mental states to dispositional or behaviouristic states (Luckhardt, 1983: 319). Given the pervasiveness of this reading, Wittgenstein becomes a relevant force to reckon with when considering the mind-body debate.

The "core" of the PLA is stated in §§243ff. of the *Investigations*. As such, the theses put forth there will be the main focus of our investigation. I aim to explore whether the PLA, strictly speaking, is deleterious to mind-body dualism, not whether Wittgenstein would approve of a dualistic philosophy of mind. That being said, the PLA is notoriously ambiguous if read on its own, which means that references to other parts of the *Investigations* will be inevitable, if only to elucidate what is actually meant by it.

In order to explore, then, whether we can support any kind of body-mind dualism and accept, at the same time, the validity of the PLA, I will proceed as follows. I will begin by challenging some of the so-called "orthodox interpretations" (Stern, 2011: 331) of the PLA, and proposing what I find to be its most plausible characterization, namely, the PLA as a special case of a general problem with identity and ostension. After that, I will dedicate some space to addressing specifically behaviouristic concerns regarding the PLA. Finally, I will discuss the different ways in which we may understand

mind-body dualism and I will show that the PLA does have a bearing on some sorts of dualism, while not necessarily on others.

II. Elucidating the PLA

As I have mentioned, the PLA is notably obscure, and figuring out what it actually means has preoccupied much of the literature on the topic. Wittgenstein's aphoristic style has not lent itself to easy formalization, so much so that some have argued that interpreting §§243ff. as an "argument" of any sorts is to misconstrue Wittgenstein's point (Stern, 2011: 342-3). Wittgenstein, according to this line of reasoning, would not be interested in "proving" the impossibility of a private language by means of a *reduction ad absurdum*; rather, his aphoristic style would be warranted by the need to *show*, not prove, the unintelligibility of the thesis. This interpretation is not without its merits, and it is probably adequate to the latter Wittgenstein's general anti-theorizing attitude (Pears, 1988: 214-215).

Nevertheless, I think it will be most appropriate for us to treat it as an argument, if, perhaps, not as a simple *reductio ad absurdum*, due to several reasons. First of all, some level of formalization is useful if we are to objectively assess the implications of Wittgenstein's treatment of private languages; we will hardly be able to draw clear conclusions from vague aphorisms taken at face value. Aside from that, however, since our aim relates to the PLA, and to how it has been covered by the literature, it does not seem necessary to dialogue with Wittgenstein's idiosyncrasies if they do not directly contribute to this particular academic debate. Thus, I will attempt to formalize the PLA, beginning by showing why I think two prevalent approaches – the "fallibility of memory" approach and the "verificationist" approach – are lacking. Subsequently, I will present my own interpretation.

a. The PLA as memory scepticism

Most commentators who take the PLA as an argument proper seem to agree, at least, that it implies what follows:[106]

(1) A private language stands in opposition to a public language insofar as the meaning of its terms is privately set.

(2) No meaning can be privately set.

(3) Therefore, there can be no private language.

Many of these terms are in dire need of definition, and that is so intentionally, since the crux of the dispute resides in how we come to understand them. Namely, we will see that what we mean by "privately set", and what we base (2) on, will suppose the main source of disagreement among commentators. I will call those nodes of dissent the "privacy clause" (PC) and the "criterion-setting clause" (CC), which should be added as elided premises to the main argument.

The "fallibility of memory" approach, most famously defended by A. J. Ayer, interprets the PLA to entail this:

> (PC): A language is private when the objects it refers to are, themselves, private. (Ayer, 1954: 64)[107]
>
> (CC): No meaning can be privately set because, if we grant that we cannot immediately ascertain how to use a private term, then we cannot trust any of our private grounds for evidence. (*ibid*.: 68)[108]

[106] For a general review of the bibliography around the PLA, *vid.* (Villanueva, 1975a), (Villanueva, 1975b) and (Stern, 2011).

[107] "What philosophers usually seem to have in mind when they speak of a private language is one that is, in their view, necessarily private, in as much as it is used by some particular person to refer only to his own private experiences".

[108] "For if one cannot be trusted to recognize one [private sensation], neither can one be trusted to recognize the other".

Once the PLA is set up like this, Ayer has strong reasons to dismiss it. If we cannot trust any of our private grounds of evidence (sight, memory, etc.), then it is not just private languages that cause trouble; it appears to be impossible to use *any* language (*ibid.*).

Much has been written against Ayer's interpretation on both fronts. Regarding (PC), Ayer has been accused of misinterpreting what is relevant about the hypothetical private language that Wittgenstein discusses. It is not that it denotes private objects, but that it is a language that *nobody but its user may, even potentially, come to learn*.[109] (Thomson, 1964: 20; Villanueva, 1975: 81; Luckhardt, 1983: 327; Stern, 2011: 333). We are detaching ourselves from the realm of the languages that we *actually* do use (Candlish, 1980: 86) since, as far as we know, our languages are inter-translatable, and, what is more, the entire point of the PLA is to show that any such languages are a logical impossibility. Ayer's (CC) does not fare much better. I will expound more on this point, but Wittgenstein's concern is not that we may be "fallible" when confined to our private fora; rather, that there cannot be anything like a criterion of correctness, fallible or not, that is entirely private (Pears, 1988: 333). In other words, Ayer interprets (CC) to mean something akin to: there is a process *P* by which I identify private objects and I name them. *P* has a non-zero chance of failing, therefore, *P* is not to be trusted. The actual clause in the PLA seems to be, on the contrary, that there can be no such process *P*.[110]

b. The PLA as verificationism

Some of our points merit further elucidation, and they will be subject to closer examination in the next section. Before doing that, however, it is necessary that we

[109] This is evidenced, for instance, in PI §261. "'Has' and 'something' also belong to our common language". What does he mean by this? Our putative private linguist is one which *in no way* depends on terms whose meaning may be publicly communicated; it is a matter of fact that we can publicly talk *about* "private objects" – as evidenced by this very paragraph. That does not mean, however, that we can "denote" private objects.

[110] Wittgenstein does not merely say that we *can fail*; "in the present case, I have no criterion of correctness" (PI §258)

address the interpretation Judith Jarvis Thomson proposes for the PLA. She contends that we ought to interpret said clauses as follows:

> (PC): A language is private when it is logically impossible for anyone but its sole user to understand it. (Thomson, 1964: 21)
>
> (CC): No meaning can be privately set, because, for a sign to be a kind-name, it must be possible to *find out* (publicly) whether a thing is of that kind, which amounts to the principle of verification (*ibid*.: 29).

Thomson's assessment of the (PC) seems entirely adequate, but I cannot agree with her understanding of the (CC) as a restatement of the verification principle. The reason why will be made clearer in the following section, but, as of now, we can consider an example, and compare Wittgenstein's actual assessment to a "verificationist" one.

Compare a person undergoing tremendous pain to a great actor, who mimics "pain-behaviour" to such degree of perfection that there is no discernible difference between their acting and actual pain-behaviour. Consider this actor performing such moving scene on a stage. We can *tell* that this person is not "actually" in pain, they simply excel at their art, and, if pressed, we can mention other extraneous factors to support our judgment, like them not leaving the stage, the normal reactions of their colleagues, and so forth. But how does that amount to *finding out*, in a verificationist sense, that they are not, in fact, in pain? Is it essential to pain-behaviour that it transpires outside a stage? Or that witnesses react in a given manner?[111] It does not appear so, yet we do not think it misjudged to say that the actor was not *actually* in pain, and, more

[111] The main question here revolves around the possibility of mimicry, deception or acting. There are two theses that seem to conform to our common understanding. (1) Somebody may convincingly fake the behaviour associated to a mental state; (2) faking a behaviour implies not being in the mental state typically associated to it. Given that, there are contexts where there is an expectation for faking, and where it seems we can be said to, despite perceiving the exact same behaviour that would make us think that somebody is experiencing some mental state, *tell it apart* from the "actual thing". If this is so, a purely verificationist stance does not adequately portray the language games at stake here, which go "beyond" denotation. (Cf. Putnam, 1980: 29)

importantly, neither does Wittgenstein.[112] The verificationist approach does not seem to be compatible with Wittgenstein's actual theses.

c. The general problem with ostension

In order to present our interpretation of the PLA, we need to characterize a general issue present in Wittgenstein's philosophy, of which the PLA would be a particular case. This issue is brought up by Thomson in the previously cited article:

> The question, "How do I identify a kind of sensation?" is a very respectable philosophical question. But of course it is only a special case of the very respectable philosophical question, "How do I identify a kind of *thing*?" (*ibid*.: 26)

And I cannot but agree with her, since therein lies the question. The *Investigations* may be primarily concerned with so-called "inner states", but it starts off as a general discussion about language use. How do we come to use any word whatsoever, if, for any rule on word-using we may encounter, we would need yet another rule on rule-following, falling into a *regressus ad infinitum*? (*vid. PI* §86; Kripke, 1984: 62)

This problem has been, perhaps, most famously exposed by Saul Kripke in his work on the PLA. I will skim over his very suggestive interpretation, due to space constraints. Shortly, Kripke posits that the *Investigations* are chiefly concerned with a so-called "sceptic" objection to all rule-following (*ibid*.: 8). It seems as though we act in certain ways – for instance, giving the "correct" answers to arithmetic problems – because we follow specific rules in doing so. Nonetheless, we can never be confident that we are justified in following a rule, because any amount of past instances that, putatively,

---

[112] "'But you will surely admit that there is a difference between pain-behaviour accompanied by pain and pain-behaviour without any pain?'–Admit it? What greater difference could there be?" (PI §304)

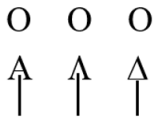
express the rule, could be made fit into arbitrarily many other rules. A finite set of past instances is never "enough evidence" for a particular general rule, since there can always be counterexamples not covered by that finite set that we add to construct an apropos alternative rule. (*ibid*.: 18).

I have no objections to this interpretation; I am, in fact, highly indebted to it. Nonetheless, we need not operate at such a level of abstraction to make our case. I will apply Kripke's problem of rule-following to the more concrete problems of ostension and identity, which will be shown to be necessarily intertwined.

In order for us to "point at things" and use these crude denotations as the building blocks of a language, as does Wittgenstein in PI §2, we surely need to have a criterion to distinguish what is "equal" from what is not. That is, in order for me to be proficient at bringing bricks, I need to know which two things are equal *qua* bricks, or, what "about the brick" is being pointed at when I am taught what a brick is[113]. If I do not, I will not be able to follow the task at all. Let us move away from bricks, and consider the Greek alphabet. Alpha, delta, and lambda are three similar-looking yet completely different letters; my proficiency at decoding the name:

ΑΛΚΙΒΙΑΔΗΣ

depends on my being able to understand which differences between letters are "significant" to this task, namely, reading a Greek name. The first and second alphas are not "the same", insofar as they are two distinct "tokens", but they are of the same type. Now consider the following diagram:



[113] Cf. *PI* §33. Am I pointing to the rectangular shape of the brick's face? To its colour? Ostension is always ambiguous on its own.

It is not clear now whether we are before three tokens of the same type or not. They seem to be slightly differing arrows, which point at circles nevertheless. They have been deprived from their original context, and what constitutes "equalness" in this new environment becomes ambiguous.

All of this is to illustrate Wittgenstein's note in PI §215: "Then are two things the same when they are what *one* thing is? And how am I to apply what the *one* thing shews me to the case of two things?" Equalness itself, which seems to be at the heart of all possible ostension, requires a criterion, because it is not self-evident (Pears, 1988: 386). Let us turn back to the experiment described in §258. Our private linguist has some sensation on day 1, let us call it S_1, accordingly, and she denotes it with an *S* on her diary. It is day 2 now, and she experiences S_2. How is she to proceed now, to know whether it is appropriate to write down an *S* or not? *Contra* Ayer, we can endow our linguist with a perfect memory, she can mentally reconstruct her every past state at will. Yet, she will have to compare S_1 and S_2 and emit a judgment on whether they are tokens of the same type or not. How will she do it? Trying to appeal to higher-order disambiguation criteria just moves the problem a step backwards, since she will have to wonder whether this situation is "equal" to a past situation where disambiguation rule *R* applied, running again into the *regressus* problem.

We seem to have reached an impasse, since this objection would apparently hold equal to private and public languages. There is, nonetheless, a crucial distinction. The public forum establishes what we may call *semi-rigid grounds of significance*, which are to be regarded as "brute facts" given that we *can* use language to communicate (Kripke, 1984: 98). (1) There needs to be some regularity in the world in order for us to be able to assign meaning to our terms. If things could never reliably be said to have a colour, and our visual perceptions were chimeras, we could hardly be expected to come

to apprehend a colour-language.[114] (Rhees, 1954: 93). The "stubbornness" of reality makes quaddition-like formulas ultimately unusable. But, of course, what counts as "regularity", as we have already discussed, cannot be established *a priori* (Kripke, 1984: 105); we also need (2) some regularity in the linguistic uses of the community of speakers, to whom the appropriate use of terms is of some significance, and are thus able to enforce it and teach it. (Kripke, 1984: 96; Pears, 1988: 370).

Why cannot we disambiguate ostension by purely private means? A private context runs into regression problems because there is never a "last" ground of justification that is not "simply chosen" to be so, and thus amounts to not distinguishing between "being right" and "feeling right", which is the entire point (PI §258). Meanwhile, the tree you run into, or the teacher that corrects you, do not admit further appeals, they are "coercive" in their disambiguation.[115] This does not mean that, analysed in the abstract, public practices are without ambiguity, but the coerciveness of use overrides the need for an "indubitable" grounding.

The world and our language colleagues conform the necessary context in which we can disambiguate ostension, they are "what happened before and after the pointing" (PI §35). They are only *semi-rigid* grounds, because the meaning of our words does change, and there is room for idiolectal variation, but this has to be ultimately constrained within the bounds of usability. We only get to disambiguate our terms if there is any consequence to getting them wrong. Therefore, after this laborious exercise at elucidation, I can give my proposed interpretations of (PC) and (CC):

---

[114] I interpret PI §80 to serve a double function. On the one hand, at face value, it is a reflection about how our rule-following does not depend on our effectively being able to know how to use the rule under outlandish circumstances. But, additionally, it points out how, for a rule to be meaningful, there needs to be some regularity to the cases where it applies. Our language about chairs is not equipped to talk about flickering and disappearing objects, because it does not need to be. If chairs did flicker and disappear, however, our language would not be appropriate.

[115] Cf. *PI* §303. "Just try–in a real case–to doubt someone else's fear or pain".

(PC): A language is private when it is logically impossible for anyone but its sole user to understand it.

(CC): No meaning can be privately set because any criterion of identity, public or private, requires disambiguation through semi-rigid grounds of significance. Any purely private context fails in this regard because it is always subject to a *regressus ad infinitum* where no ultimate ground of justification is to be found. Public contexts find a way to halt the regression by coercing the speaker through the necessity to act.

III. Dualism and the PLA

Having elucidated an operational form of the PLA, we can now move on to analysing our main concern, whether it has any bearing on mind-body dualism. I will begin by addressing a possible way to interpret the PLA that would immediately discard any kind of dualism.

a. The behaviourist challenge

We have established that, for any term to have a set meaning, it requires a public context of disambiguation, making all attempts at constructing a private language meaningless. What does this entail, however, for the terms we allegedly use to denote private objects like "pain"? A behaviourist interpretation seems simple enough:

(1) Terms like "pain" have meaning for us.

(2) Per the PLA, no meaning can be privately set.

(3) Therefore, the meaning of "pain" is publicly set.

(4) Pain-behaviour is public, pain-sensations are private.

(5) Therefore, the meaning of "pain" cannot be grounded on pain-sensations.

(6) Therefore, "pain" denotes pain-behaviour.

If this is so, we do not use "pain" to denote anything about our inner experience, but about some behaviour through which we can be taught to talk about "pain", and through which a certain use of the word can be enforced. Wittgenstein, in fact, is adamant about how the word "pain" is not used to *denote* or *describe* any hidden mental state (PI §290). Is all talk about dualism linguistic nonsense then, dissolved by the PLA?

Not quite. There is a clearly unjustified leap from (5) to (6). Up until then, the argument holds, it is true that under the provisos of the PLA we cannot ground the meaning of "pain" on pain-sensations (Pears, 1988: 350). But that does not mean that "pain" needs to have a *denotative* function of any sort, let alone that it needs to denote *behaviour*. Recall our earlier example about the actor. How do we explain it under our current interpretation of the PLA? As members of a community, we come to disambiguate our references to pain on contextual bases. We can only tell whether certain behaviour is "pain", or "acting", or anything at all based on how our linguistic community has acted regarding certain scenarios and how they have enforced the use of certain rules. We can imagine a child going for the first time to a theatre and telling his parents that the actor needs help, an assessment that the parents would correct by noting how "pain" – the "pain-language game" – does not apply in that situation.[116]

There is a difference between assertability conditions (Kripke, 1984: 111) and the meaning of a term. A given behaviour is necessary for us to come to learn the meaning of a sensation-term, but the meaning of a sensation-term is not exhausted by behaviour (vid. Putnam, 1967: 57-8; Luckhardt, 1983: 328).

b. What do we understand by dualism?

What we can derive from the past discussion is that, even if we grant that there are such things as "private objects", we would not be able to define them by ostension without the

[116] Cf. *PI* §584.

concurrence of publicly enforced criteria. Our discussion about the private linguist distinguishing between S_1 and S_2 tacitly implied their persistence as "objects" of some sort, and their ontological status played no role in our argument. Strictly speaking, then, the PLA says nothing about the ontology of the private forum, and, in a trivial sense, it is compatible with any manner of dualism.

But this answer is not satisfactory, because, as J.J.C. Smart puts it, even though a state of affairs about our mental reality may be compatible with several explanations, mere compatibility is not enough to merit accepting any one of them (Smart, 1959: 155-6). The question we should be asking is, do we have any reasons to maintain dualism given the PLA?

Mind-body dualism comes in many different shapes. A tripartite distinction that has enjoyed some popularity, regarding the different ontological presuppositions that dualism may have, is that of substance, property, and predicate dualism (Robinson, 2020), in decreasing order of ontological commitment. We may characterize them as follows:[117]

Substance dualism, which would be a thesis such as the one espoused by René Descartes, holds that:

(1) There are mental states, different from physical states.

(2) The mental states of a subject S correspond to mental properties of S.

(3) These mental properties belong to a distinct mental substance.

[117] This is a decidedly simplistic account of what are deeply complex theories about the mind. Nonetheless, focusing on these three particular theses seems to (a) show the main sources of disagreement between the three positions, (b) establish some points whose contention against the PLA seems most relevant. Property dualism, for instance, may say much more than the very vague theses (1) and (2) would have it, but it is of the utmost importance whether the PLA posits serious problems to those theses; more so than other, perhaps, less central stances within such theory.

Property dualism does not commit to thesis (3), being compatible with the idea that there are only physical substances; emergentism being an example of it (Mitchell, 2010: 172)[118], and predicate dualism does not commit to either theses (2) or (3), basing itself, for instance, on multiple realizability to argue that mental states, while reducible as *tokens* to physical states, are not so reducible as *types*; an example of this being Davidson's anomalous monism (Davidson, 1970: 99-100).

By characterizing these three classes of dualism, we can see that the more ontologically compromised theories necessarily entail the less compromised ones. Substance dualism, for instance, as characterized, would entail both property and predicate dualism. There could be other stances, but these seem to be the most useful for our discussion.

Let us begin, then, by assessing the plausibility of thesis (3) vis-à-vis the PLA. The PLA says nothing about ontology, but what reasons could we have to support the existence of a mental substance? Descartes argues that, since we first come to be certain of our being mental, and we can have a clear and distinct notion of it (2011 [1641]: 76-7), the union between our mental and physical states is to be held as contingent, and, thus, said states correspond to different substances.

If we accept the PLA, however, we cannot sustain that there be any privilege in acquiring knowledge about our mental states; Cartesian introspection is required to make this argument work, but the PLA forces mental concepts to be set in the very same public forum as physical concepts, as we have already seen. Thus, the source of distinctiveness that Descartes alleges as sufficient reason to defend the existence of a *res cogitans* is lost. The PLA equalizes the ground for all states, mental and physical, there is no priority

[118] Mitchell tackles emergence *tout court*, but that is, naturally, applicable to our case.

other than that established by the linguistic uses of the community. Therefore, multiplying the substances, if we are to accept the PLA, seems capricious.[119]

What about thesis (2)? The argument for characterizing mental states as properties of their own can take many shapes. We can consider emergent states, says Mitchell, as relevant entities subject to natural selection, for example, or as causally efficient (2010: 179-80). This seems like the sort of hypostasising that Wittgenstein would forbid (*vid.* PI II§76), but we must recall that we are not accepting or assessing all of Wittgenstein's psychology, we are merely addressing the relation between the PLA and dualism. Does the public setting of meaning affect in any way the assessment that mental states may be considered as properties of physical substances, insofar as they are causally efficient, or insofar as they are subjected, by their own, to natural selection? There may be other arguments against them, but it does not seem that what is posited by the PLA alone does anything to problematize them. We learn publicly, for example, to refer to some mental state of ours that precedes our *acting* as "determination". We cannot learn to use the word by ourselves, but this does not preclude that, once we learn to use it, we think it best to analyse it as a property that instantiates onto us.

If this is so, and the PLA does not pose serious problems to thesis (2), *a fortiori* it will not be problematic for thesis (1). Thesis (1) is not directly implied by thesis (2), since it is an assertion about there *being* mental states. But this one has been already tackled by our previous discussions regarding verificationism and behaviourism. Privacy is not discarded from our language games. Our using it and talking about it cannot be fully independent from the physical and public (Pears, 1988: 350), but that does not entail that we may *reduce* mental states to physical states (Luckhardt, 1983: 329). Terms about

[119] Cf. PI §293; the example of the beetle in the box is particularly relevant to this effect. Once again, it is impossible to tackle all forms of substance dualism. The main argument, in any case, is that a particularly strong form of first person privilege seems necessary to deem substance dualism a reasonable position to sustain. If someone were to convincingly sustain the necessity for two *substances* even without said privilege, then it would not fail the PLA test either.

mental states are, as a matter of fact, present all throughout our common linguistic experience, and they do not necessarily seem to be reducible.

IV. Conclusion

It may be legitimately objected that the sort of dualism that withstands the PLA is so far removed from Cartesian dualism that it is not appropriate to even consider it at the same level. However, it does not appear that the discussion has been fruitless. I have attempted to present a non-behaviourist, non-verificationist view of the PLA, which allows for far more flexibility in the status we may attribute to mental states. If I have succeeded in my argument, the PLA does not reduce mental terms to their behavioural counterparts, it simply establishes general conditions for criteria-setting, which then may apply beyond the strict scope of what is publicly verifiable. This is a notable shift from the starting point, and, if it is not to be called dualism – although emergentist stances are typically called dualist (Gregory and Zangwill, 1987: 204) – it certainly is not pure physicalism.

There are many interesting topics relating to the general PLA discussion we have not been able to tackle here, and which may warrant further research. Is reductive physicalism even intelligible from a Wittgensteinian point of view since it ignores the problems regarding rule-following and criteria of identity? What are the links between neutral monism and language games grounded on an indeterminate sort of world regularity? Can there be an ontologically uncompromised Wittgensteinian functionalism? How does the PLA fare with qualia? These, among many others, are questions we will have to leave unanswered for the time being.

Works cited

Ayer, A. J. and Rhees, R. “Symposium: Can There Be a Private Language?”, *Proceedings of the Aristotelian Society, Supplementary Volumes,* 1954, 28, *Belief and Will* (1954), pp. 63-94.

Candlish, Stewart. “The Real Private Language Argument”, *Philosophy* 55(211) (Jan., 1980), pp. 85-94.

Davidson, Donald. “Mental Events”, in *Experience and Theory*, Lawrence Foster and J. W. Swanson (eds.). University of Massachusetts Press, 1970.

Descartes, René. *Meditaciones Metafísicas*. Transl. Guillermo Graíño Ferrer. Madrid: Alianza Editorial, 2011 [1641].

Kripke, Saul A. *Wittgenstein on Rules and Private Language*. Cambridge: Harvard University Press, 1982.

Gregory, Richard L. and Zangwill, O. L. (eds.). *The Oxford Companion to the Mind*. New York: Oxford University Press, 1997.

Luckhardt, C. Grant. “Wittgenstein and Behaviorism”, *Synthese* 56(3) *Ludwig Wittgenstein: Proceedings of a conference Sponsored by the Austrian Institute, New York, Part II*, (Sep., 1983), pp. 319-338.

Mitchell, Sandra D. “Emergence: logical, functional and dynamical”, *Synthese* 185, (2012), pp. 171-186.

Pears, David, *The False Prison. Study of the Development of Wittgenstein’s Philosophy. Volume Two.* Oxford: Oxford University Press, 1988.

Putnam, Hilary. “The Nature of Mental States”, in W. H. Capitan y D.D. Merrill (eds.), *Art, Mind and Religion*. Pittsburgh University Press. 1967

- “Brains and Behavior”, in Ned Block (ed.), *Readings in Philosophy of Psychology. Volume One.* Harvard University Press. 1980

Robinson, Howard, "Dualism", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/ fall2020/ entries/dualism/>.

Smart, J. J. C. “Sensations and Brain Processes”, *The Philosophical Review*, 68(2) (Apr., 1959), pp. 141-156.

Stern, David. “Private Language”, in *The Oxford Handbook of Wittgenstein*, Oskari Kuusela and Marie McGinn (eds.). Oxford: Oxford University Press, 2011.

Thomson, Judith Jarvis. “Private Languages”, *American Philosophical Quarterly*, 1(1) (Jan., 1964), pp. 20-31.

Villanueva, Enrique. “El argumento del lenguaje privado (I)”. *Crítica: Revista Hispanoamericana de Filosofía*, 7(20) (Oct., 1975), pp. 73-104.

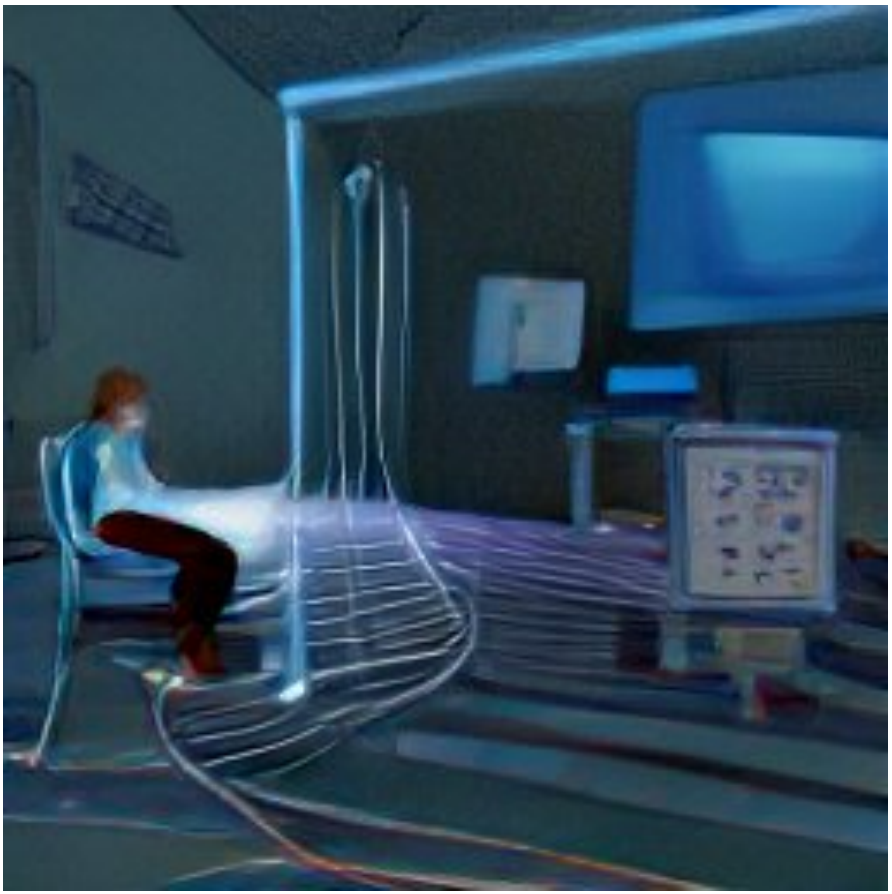
- “El argumento del lenguaje privado (II)”. *Crítica: Revista Hispanoamericana de Filosofía*, 7(21) (Dec., 1975), pp. 18-33.

Wittgenstein, Ludwig. *Philosophical Investigations* (PI). Transl. G. E. M. Anscombe. Oxford: Basil Blackwell Ltd. 1967 [1953].

- *Investigaciones Filosóficas*. Transl. Jesús Padilla Gálvez. Madrid: Trotta Editorial. 2017 [1953].

Álvaro is a last year Philosophy student at Universidad Nacional de Educación a Distancia (Spain), and he recently received his BA in Philosophy, Politics and Economics at Universidad Carlos III de Madrid. His research interests include Kant's philosophy and the philosophies of mind, action and biology. Starting in October 2022, as part of "La Caixa" fellowship for postgraduate studies abroad, he will be studying a Masters by Research in Philosophy at Durham University (UK). Aside from philosophy, Álvaro likes to participate at radio shows, gardening, and (poorly) playing chess.

Biology or Information: Refuting the Simulation Argument



Brady Cook

Abstract

In his seminal 2003 paper (*Are You Living in a Computer Simulation?*), Nick Bostrom argues that provided one accepts a few basic assumptions, one must also accept that our Universe is almost certainly a *simulation*. I will show that *if* his argument is used to draw an epistemic claim, then it is reducible to absurdity; and *if* it is used to draw an ontological claim, then it relies on an unjustified – and implausible – presupposition. Lastly, a conceptual error within the arguments mathematical model will be uncovered. These considerations will show, and this paper will thus conclude, that the widely supported *simulation argument* is false.

All of this [pointing to the stars] might just be an elaborate simulation running inside a little device sitting on someone's table.

– Capt. Jean-Luc Picard (Patrick Stewart).

I. Introduction

Many believe that me, you, and our entire Universe – including every thought and experience – supervenes over some complex computer simulation. This hypothesis has been pondered by physicists for several decades, but has seen unprecedented interest following a recent development: the publication of Nick Bostrom's *simulation argument* (*SA*). A probabilistic analysis ratiocinated across many disciplines, and, in its reductionist form, amongst the general public as well. While its widespread recognition is partially due to pundits frequently voicing their support (e.g., Neil deGrasse Tyson, etc.), there is another reason behind *SA* garnering so much attention. That is, if sound, it derives a remarkable conclusion about the implementation of our Universe; and does so from rather simple empirical assumptions. It is rare to gain so much leverage out of a short philosophical argument.[120] Before explicating its reasoning, however, I must first explain what this project is *about*. I will show that – depending on how it is employed (epistemically or ontologically) – the simulation argument can be either reduced to *absurdity*, or shown to rely on an implausible presupposition. Moreover, I will also uncover a conceptual error within the argument's mathematical model.

To be clear, I will *not* show that the Universe is biological. Nor will I show that it has a significant probability of being so. I *will*, however, provide an explication of Bostrom's argument (and its implications) while calling attention to several flaws, each of which invalidate its conclusion. This, in turn, will undermine the leading justification for a non-biological Universe.

---

[120] Nick Bostrom, *FAQ Section*, https://www.simulation-argument.com/faq.html.

II. The Simulation Argument (SA)

Most technologists believe that there will be enormous amounts of computational power available in the future. Enough to simulate the entire history of our Universe *many* times over. If this is correct, these future civilizations may run highly detailed simulations of their forbears (or people like their forebears), and because their computers would be so powerful, they could run a great number of simulations.[121]

Now consider that the forebears in these simulations are conscious like us. It would then follow that the vast majority of observers that will exist (with experiences like ours) will be simulated rather than biological. If this were the case, Bostrom (2003) argues that we would be rational to think that we are likely among the simulated minds rather than among the biological ones.

> If we don't think that we are currently living in a computer simulation, we are not entitled to believe that we will have descendants who will run lots of such simulations of their forebears.[122]

This is the gist of his argument; however, he offers a *formal version*, claiming that at least one of the following propositions is true:

1. The human species is very likely to go extinct before reaching a posthuman stage.[123]
2. Any posthuman civilization is extremely unlikely to run a significant number of simulations of their evolutionary history (or variations thereof).

---

[121] Nick Bostrom, *Are we Living in a Computer Simulation?* (2003), p. 2.

[122] Bostrom, p. 2.

[123] A *posthuman stage* refers to a period of human evolution in which technological capabilities are vastly superior to what we (today) would consider 'human'. In this context, it denotes a period in which civilizations are capable of simulating vast numbers of conscious beings whose experiences are indiscernible from our own.

3. We are almost certainly living in a computer simulation.[124]

If this tripartite disjunction is true, one of its propositions *must* be true. Which would imply that, if one denies the first two propositions, and they are indeed acting rationally, then they must commit themselves to the truth of the third (i.e., that we are almost certainly living in a computer simulation). There are two *assumptions*, however, which Bostrom needs to establish this disjunctive claim.

Assumption-A is common in the philosophy of mind; that is, the *substrate-independence thesis*. This asserts that "mental states can supervene on any of a broad class of physical substrates."[125] In other words, if a system implements the right sort of computational structures and processes, it can be associated with conscious experience. It is not an essential property of consciousness that it is implemented on carbon-based biological neural networks. This assumption is necessary for if consciousness relies on biological substrates, then it cannot be *simulated* (*1* would thus be true and *3* would be false). According to Bostrom, however, the substrate-independence thesis is widely accepted among cognitive scientists and philosophers of mind.

Assumption-B regards the technological limits of computation. Specifically, it is required that posthuman civilizations have enough available computing power to perform a sufficiently large number of simulations. Citing the work of several technologists and computer scientists, Bostrom considers $\sim 10^{33}$ - 10^{36} operations per second to be a fair estimate of the computational power necessary to perform a realistic simulation of the entire mental history of humankind. Then, citing the work of R.J. Bradbury, he notes that a computer

[124] Bostrom, p. 1.

[125] Bostrom, p. 3.

powered by a *Dyson sphere*[126] (with nanotechnological designs of the early 2000's) could perform an estimated 10^{42} operations per second. Given this, Bostrom thinks it is safe to assume that posthuman civilizations would have enough computing power to run an astronomical number of ancestor-simulations, even while using only a tiny fraction of their resources for that purpose.[127]

With these assumptions in mind, we can now get to the crux of Bostrom's argument. He makes use of some formal probability here, which I will reproduce verbatim before offering an alphabetic translation. Let us start by considering the following notation (quoted directly from Bostrom, 2003):

> f_p: Fraction of all human-level technological civilizations that survive to reach a posthuman stage.
>
> $\overline{N}$: Average number of ancestor-simulations run by a posthuman civilization.
>
> $\overline{H}$: Average number of individuals that have lived in a civilization before it reaches a posthuman stage.
>
> The actual fraction of all observers with human-type experiences that live in simulations is then:
>
> $$f_{sim} = \frac{f_p \overline{N}\,\overline{H}}{(f_p \overline{N}\,\overline{H}) + \overline{H}}$$
>
> Writing f_i for the fraction of posthuman civilizations that are interested in running ancestor-simulations (or that contain at least some individuals who are interested in that and have sufficient resources to run a significant number of

[126] A *Dyson sphere* is a hypothetical artificial structure capable of capturing large percentages of a star's power output.

[127] Bostrom, p. 7.

such simulations), and $\overline{N_I}$ for the average number of ancestor-simulations run by such interested civilizations, we have:

$$\overline{N} = f_I \overline{N_I}$$

And thus:

$$f_{sim} = \frac{f_P f_I \overline{N_I}}{(f_P f_I \overline{N_I}) + 1} \qquad (*)$$

Because of the immense computing power of posthuman civilizations, $\overline{N_I}$ is extremely large. By inspecting (*) we can then see that *at least one* of the following three propositions must be true:

(*1*) $f_P \approx 0$

(*2*) $f_I \approx 0$

(*3*) $f_{sim} \approx 1$

More generally, if we knew that a fraction *x* of all observers with human-type experiences live in simulations, and we don't have any information to indicate that our own particular experiences are any more or less likely than other human-type experiences to have been implemented *in vivo* rather than *in machina*, then our credence that we are in a simulation should equal *x*:

$$Cr(SIM \mid f_{sim} = x) = x \qquad (\#)$$[128]

I will now offer an alphabetic translation of the probability theory just presented. Bostrom first estimates the fraction of all people in existence that are simulated (*fsim*). This is the expectation of the number of simulated people divided by the expectation of the number of simulated people plus the number of non-simulated people. Note, the expectation of the number of simulated people is equal to the probability of simulations being done times the average number of simulations that would be done (if simulations were done) times the average number of people in each simulation.[129]

Translating this fraction into slightly different notation, it follows that – because the number of simulations run by a civilization capable of running them would be very great (*Assumption-B*) – unless there is a very low fraction of simulations being done (practically null), then there is an extremely high fraction of simulated people in existence (practically unity).

From here, Bostrom makes an appeal to the *principle of bland indifference* – a non-informative prior adopted from Bayesian statistics.[130] Essentially, the principle (hereinafter referred to as PBI) states that if there are x possible outcomes and there is no reason to view one as being any more likely than another, then each should be assigned a probability of 1/x. For example, if we are flipping a fair coin, then the odds assigned to landing either side (heads vs. tails) should be 1/2. The principle of bland indifference holds. However, if we learn that the coin is weighted to land on heads, then the odds assigned should no longer be 1/2. The principle of bland indifference no longer holds.

---

[128] Bostrom, p. 7-9.

[129] This paragraph has been (loosely) extracted from Brian Eggleston's *Review of Bostrom's Simulation Argument* (undated).

[130] A Bayesian *prior* is a probability distribution that would express one's beliefs about some quantity before some evidence is taken into account.

Applied to Bostrom's argument, PBI tells us that the probability of living in a simulated Universe instead of a biological one should be considered equal to the fraction already established ($fsim$). That is because, as it stands, we have no evidence to suggest that our own experiences are more or less likely than other human-type experiences to be biological rather than simulated.

Notice then, if future civilizations are expected to run a significant number of simulations (which would value the fraction of simulated people in existence [$fsim$] at almost unity), and PBI is applied, then the fraction that we ourselves live in a simulation is the same (almost unity). This demonstrates that one of two things must be true: *either we are living in a simulation, or our descendants will almost certainly never run a significant number of ancestor-simulations*.

II - i. Important Clarification

Explained just now is *the simulation argument* (*SA*), which suggests a direct relation between how likely it is that we (humans) will one day create ancestor-simulations, and how likely it is that we ourselves are in one. It also suggests an epistemic dependency between these propositions. That is, if it is *believed* that we will likely create ancestor-simulations, then it should also be *believed* that we are in one. Remember: *SA* suggests the existence of a relation between propositions, *not* that our Universe is simulated. The latter is a separate hypothesis, hereinafter denoted by *SIM*.

III. Objections

Almost all objections to the argument (SA) have attempted to refute its operative assumptions (e.g., the limitations of computational power, the substrate independence thesis, etc.).[131] Notice, however, these are not actually tackling the arguments logic. Which says *if* all stated assumptions are true *then* some fact

[131] A popular example of one such objection can be found in Jonathan Birch's *On the "Simulation Argument" and Selective Skepticism* (2013). Birch accuses Bostrom of being selectively skeptical by presupposing that we possess good evidence for claims about the physical limits of computation and yet lack good evidence for claims about our own physical constitution.

about the world is true.[132] It is a *conditional claim* which does not depend upon the assumptions actually being true, but what logically follows from their truth. It is an extremely persuasive argument (as it is yet to be refuted[133]); although in this section, I will refute it.

III – i. Conceptual Error

My first objection will display a conceptual error within Bostrom's probability theory. Let us start by reconsidering the following notation (quoted directly from Bostrom, 2003):

f_P: Fraction of all human-level technological civilizations that survive to reach a posthuman stage.

f_I: Fraction of posthuman civilizations that are interested in running ancestor-simulations (or that contain at least some individuals who are interested in that and have sufficient resources to run a significant number of such simulations).

$\overline{N_I}$: Average number of ancestor-simulations run by such interested civilizations.[134]

---

[132] According to SA, we need *five* assumptions to derive this "fact about the world" (i.e., that we are almost certainly living in a simulation). They are (1) the substrate independence thesis, (2) adequately high levels of computational power available for posthuman civilizations, (3) $\neg$ ($f_P \approx 0$), and (4) $\neg$ ($f_I \approx 0$). Then there is also the weak assumption just discussed in *section 5.3.* (i.e., that the average number of people living in the pre-posthuman phase is not astronomically greater for non-simulating civilizations than for civilizations that end up running significant numbers of ancestor-simulations).

[134] Bostrom (2003), p. 6.

Bostrom argues that the fraction of all observers with human-type experiences that live in simulations is:

[135]

$$f_{sim} = \frac{f_p f_t \overline{N_t}}{(f_p f_t \overline{N_t}) + 1}$$

Notice, the product of f_p and f_t alone gives us the fraction of human-level technological civilizations that perform ancestor-simulations. Thus, if the value applied to $f_p f_t$ is say 1/50, it would follow that – in order for it to be most probable that a *single* ancestor-simulation is performed – there must be at least twenty-five other human-level technological civilizations in the Universe (including past, present, and future) aside from humans. But what if someone does not believe that there are? Well, because $\overline{N_t}$ would be very high, Bostrom's model above tells them that they must still believe – indeed, on the basis of epistemic consistency – that some civilization(s) will in fact perform ancestor-simulations. Clearly, this is a problem; there must be an error in the model.

To demonstrate, consider the following scenario. Some agent X believes that *humans are the only human-level technological civilization in the Universe (including past, present, and future)*. Such a belief may be motivated by theology, a desire for significance, or abstract reasoning (e.g., the Fermi paradox). Nevertheless, the cause of the belief is irrelevant.

It would then follow that, to calculate the odds that X must accept regarding *fsim*, such that he can avoid epistemic inconsistency, $f_p f_t$ cannot be *in* the fraction, as it is in Bostrom's model, but rather extracted (both from the numerator and the denominator) and used as an *upper bound* on the value of the remaining fraction.[136]

---

[135] Bostrom (2003), p. 7.

[136] The remaining fraction would represent *fsim* assuming human-level technological civilizations become posthuman and are interested in running ancestor-simulations.

$f_P f_I$ must be multiplied by: $\overline{N_I}$ / ($\overline{N_I}$ + 1).

So in X's case, $Fsim = (f_P f_I) \times [\overline{N_I} / (\overline{N_I} + 1)]$

That is because, if X applies a probability of say 1/4 to $f_P f_I$, he is suggesting that the probability of any and all simulations existing can be no higher than twenty-five percent, given that $f_P f_I$ in his case, must apply to a *single* civilization. Within Bostrom's model, however, this is not respected. His model suggests that if X applies a value of 1/4 to $f_P f_I$, then (because $\overline{N_I}$ would be very high) X must believe that some number of simulated observers do in fact exist. Remember, however, an $f_P f_I$ value of 1/4 would not permit this belief for X, as it would suggest a majority probability (3/4) that *no* simulated observers exist, due to his anthropocentrism.

The core of the problem is that Bostrom's model *necessarily* considers all three variables (f_P, f_I, and $\overline{N_I}$) to carry equal weight in determining some supposed fraction of simulated observers. However, in X's case, they quite clearly do not, for $f_P f_I$ indicates the likelihood – supposed by X – of any and all simulations existing. So X's credence in *SIM* should be no higher than the value he applies to $f_P f_I$.

The modification I have put forth for X – extracting $f_P f_I$ and multiplying its value by the value of the *remaining fraction* (f_P / [f_P + 1]) – reflects this for f_P then affects his credence in *SIM* only within the upper bound set by $f_P f_I$.

Notice further, on this revised model, contrary to *SA*'s conclusion, X may apply a value of say 4/5 to $f_P f_I$, and thus *believe* (up to ~ 80% credence) that his descendants will perform a significant number of ancestor simulations

(¬ [*1* ∨ *2*]), without implying (*3*) that he himself almost certainly lives in a simulation. In other words, *while avoiding epistemic inconsistency, X may consider all three disjuncts to be false*. Thus, so can we the simulation argument.

With that said, you may have noticed that X's belief is not *typical*. There is no appropriate definitive standard to believe that humans are technologically unique. Rather, varying beliefs will apply. So let us consider another variable, call it $\mathbb{N}$, which respects this fact: $\mathbb{N}$ symbolizes the *number* of human-level technological civilizations in the Universe (including past, present, and future).

Notice then, $\mathbb{N}(f_P\, f_I)$ denotes the number of human-level technological civilizations in the Universe (including past, present, and future) that perform ancestor-simulations. Moreover, *iff* the value one applies to $\mathbb{N}(f_P\, f_I)$ is less than 1, then this puts them in the same predicament that X was in (*see footnote*).[137] $\mathbb{N}(f_P\, f_I)$ must then act as an upper bound on *fsim*.

$\mathbb{N}(f_P\, f_I)$ must likewise be multiplied by: $\overline{N_I}\ /\ (\overline{N_I} + 1)$.

$$Fsim\ =\ [\ \mathbb{N}(f_P\, f_I)\ ]\ \times\ [\ \overline{N_I}\ /\ (\overline{N_I} + 1)\ ]$$ [138]

For if the value applied to $\mathbb{N}(f_P\, f_I)$ is sufficiently low (e.g., 0.8), there can be no epistemic constraint to believe that the Universe is almost certainly a simulation – at least not on the basis of consistency – for one would have supposed a significant probability (1/5) that no such simulations exist. Notice, as well, that to establish a value of 0.8 for $\mathbb{N}(f_P\, f_I)$, neither f_P nor f_I need to be ≈0 (practically null). For example, each could have a value of 1/20, with $\mathbb{N}$ having a

---

[137] For in X's case, $\mathbb{N} = 1$. So for him, $f_P\, f_I$ is equivalent to $\mathbb{N}(f_P\, f_I)$.

[138] Of course, *fsim* should be calculated this way *iff* $\mathbb{N}(f_P\, f_I) < 1$. If not, Bostrom's model may suffice.

value of 320. Yet, nevertheless, it would still follow that *fsim* should not be ≈1. And so, *SA* is false. It is *not* true that one of the following propositions *must* be true:

(*1*) $f_p \approx 0$

(*2*) $f_i \approx 0$

(*3*) $f_{sim} \approx 1$

Once again, the problem is that Bostrom's model *necessarily* considers all three variables (f_p, f_i, and $\overline{N_i}$) to carry equal weight in determining how many simulated observers actually exist (without accounting for the *number* of human-level technological civilizations). However, when we account for the number of human-level technological civilizations (N), we discover certain instances – where N($f_p f_i$) < 1 – in which the three variables must *not* carry equal weight; f_p and f_i must instead carry more, working with N to generate an upper bound on *fsim*.

With that said, one may respond by claiming that individuals should not trust their beliefs regarding N. For they would first need to estimate the probability of life emerging from non-life, which is *impossible* without a second example—aside from Earth.[139] By this logic, however, we should not trust *any* of our beliefs regarding the variables in *SA*, which indicates an even weightier problem. For if we shouldn't trust our beliefs regarding the variables in *SA* then, even if we believe that *1* and *2* are false, it's no longer clear why we actually *should* believe that *3* is true. Indeed, this becomes an intractable problem.

Before advancing, I must stress that the objection presented in this section can be offered as an aside. For *SA* may be refuted while *accepting* its mathematical model. This shall be revealed by the following two objections.

---

[139] See Paul Davies', *The Eerie Silence: Renewing Our Search for Alien Intelligence* (2010).

The first of which is intended to refute Bostrom's epistemic claim (quoted directly from Bostrom, 2003):

> If we don't think that we are currently living in a computer simulation, we are not entitled to believe that we will have descendants who will run lots of such simulations of their forebears.[140]

The problem is that, when one believes that their descendants will create a *sufficient* (significantly large) number of simulations, Bostrom's claim above can be reduced to absurdity. For example, let us consider some agent, call her Z, who estimates – using Bostrom's probability theory – that the fraction of simulated people in existence (*fsim*) is a billion to one. Citing the principle of bland indifference (PBI), it is argued that Z should then take this fraction to represent the probability that she herself lives in a simulation.

Notice, however, that while Z cannot infer much information about her simulators, she can – per Bostrom's reasoning – infer that they are characteristic of her descendants, and that they have an ability, and will, to create many ancestor-simulations. It then follows that in *their* supposed reality (Z's simulators), the first two possibilities (*1* and *2*) of the tripartite disjunction are *necessarily* false. Therefore, if *SA* is valid, the third possibility (*3*) must be true. Another way of spelling this out is that, if faced with the logic of *SA*, assuming it is valid, Z's supposed simulators must accept that they are almost certainly living in a simulation. This, then, raises the question of whether Z should accept it too. I argue that she should, for she accepts Bostrom's reasoning, and that reasoning implies that her simulators – should they exist – are almost certainly living in a simulation.

From here, one may ask: what is the problem? We only seem to be bolstering the probability of *SIM* through postulating the existence of even more

---

[140] Bostrom (2003), p. 1.

simulations (or a likelihood thereof). The problem, albeit subtle, is that the logic of *SA* can be applied not just to Z and her simulators, but to her simulators' simulators as well, and so on and so forth. This is *not* an infinite regress[141] for each parent simulation should be considered progressively less likely to exist. But each by no more than the diminution from Z to her simulators given that the evidence for their existence – 1 and 2 being *necessarily* false – is stronger. Notice then, the number of simulated Universes which Z should believe to exist is *excessively* high (i.e., billions progressively stacked over a similarly large number of generations). And indeed, this is where the contradiction lies. For Z must accept that all of these simulated Universes are being carried out on a *single* computer. However, any single computer – operated by the descendants of a human-level technological civilization – will likely be incapable of performing that many highly detailed simulations, even on the most generous of expectations.

For example, Seth Lloyd of MIT has argued that if every single elementary particle in the Universe were devoted to quantum computation, it would be able to perform 10^122 operations per second on 10^92 bits of information.[142] In a stacked simulation scenario, where only 10^6 simulations are progressively stacked, after only 16 generations, the number of simulations would exceed by a factor of 10^4 the total number of bits of information available for computation in the Universe.

Even intuitively, it's a strange leap: believing that the *sum* of posthuman civilizations will perform an aggregate one-billion simulations, should not support, much less mandate, on epistemic grounds, the belief that some *single* posthuman civilization will perform immensely more – far more than *possible* by all appearances – all on a *single* device.

---

[141] See Aristotle's *Physics* (350 B.C.E).

[142] Seth Lloyd, *Programming the Universe* (2006).

The simulation argument is therefore incoherent. It merely establishes a requirement to believe that the Universe *might* be simulated – provided one rejects disjuncts 1 and 2 – to avoid epistemic inconsistency. It is contradictory, however, to suggest that unless one believes that the Universe *is* simulated, they are not entitled to believe that their descendants will run lots of such simulations themselves.

III – iii. Implausible Assumption

I will now shift focus. For *SA* seems to argue more than the *epistemic* claim just refuted; it also seems to put forth an *ontological* claim, entirely separate from belief.

> *If future civilizations are likely to perform a significant number of ancestor-simulations, then we ourselves are almost certainly living in a simulation*.

As I will show, however, this claim is highly problematic. To understand why, consider the following passage (quoted directly from Bostrom, 2003):

> If the computational cost of running even a single simulation is very great [and we are in a simulation] then we should expect our simulation to be terminated when we are about to become posthuman.[143]

This passage indicates a serious problem: *SA must assume that the type of simulations which are most likely to occur are those capable of performing nested simulations*.[144] Those with this capability will hereinafter be referred to as simulations*.

Allow me to explain. On any formulation of *SA*'s reasoning, to conclude that we almost certainly live in a simulation, there must be a prior

---

[143] Bostrom (2003), p. 7.

[144] A *nested simulation* is a simulation within a simulation.

premise stating that ancestor-simulations are likely to be performed in the future (by at least some civilization[s] in the Universe). However, if the conclusion (i.e., that we almost certainly live in a simulation) is true, but it was not true that simulations* are the most likely form of simulation, then the prior premise (i.e., that ancestor-simulations are likely to be performed in the future) should be considered false. Notice then, in order for both the premise and the conclusion to be true, it must be *assumed* that the most likely form of simulations are simulations*. Otherwise, we could move from the premise to the conclusion only by contradicting the very premise – *the conclusion would contradict its own premise*.

I suspect that this assumption, now marked, will weaken the simulation arguments appeal. In fact, the assumption may be untenable; however, let us take a closer look at what might support it. The only means of justification, I presume, would adhere to Bostrom's method of extrapolating probabilities regarding our own universe. In other words, his argument must make the further assumption that if ancestor-simulations are performed by some civilization(s) in our universe, a significant number of them will be simulations*. A *significant number* being at least however many it takes to make simulations* the most common form of simulation.

With that said, I am not confident that drawing this further assumption would work, for we may become overly presumptuous in our extrapolations. Nevertheless, let us entertain the thought for a moment as it would weaken Bostrom's argument significantly. For instance, philosopher Alexander Pruss has noted that lower quality simulations would be easier to create than higher quality simulations. Another thinker, physicist Lorenzo Pieri, has called this the *simplicity assumption* (quoted directly from Pieri, 2021):

> If we randomly select the simulation of a civilization… the likelihood of picking a given simulation is inversely correlated to the computational complexity of the simulation.[145]

In proper fashion then, we should expect "most computer simulations to be… limited in scope."[146] As I have shown, however, *SA* relies on the *majority* of simulations having great scope (i.e., having the capability to sustain multiple levels - multiple Universes). Indeed, this is a problem.

Even Bostrom himself acknowledges that "a consideration counting against the *multi-level hypothesis* [the existence of simulations*] is that the computational cost for the basement-level simulators would be very great."[147] Thus, by assuming not only that the multi-level hypothesis is true, but that it represents the *majority* of simulations that will be created, *SA* paints an implausible picture of the future.

IV. Conclusion

I must reiterate that nowhere in this work have I shown that *SIM* is false.[148] Nor have I shown a significant probability of it being so. I have shown, however, that *SA* is flawed; that it suffers from several inherent contradictions, as well as a conceptual error in its mathematical model. With that said, I am *not* disparaging the argument. It *is* incredibly powerful. Even though its conclusion (*1* ∨ *2* ∨ *3*) does not hold, it has influenced a vast range of academics – spanning many disciplines – to believe that our tangible Universe is in fact composed entirely of information. A remarkable outcome for a short philosophical argument.

---

[145] Lorenzo Pieri, *The Simplicity Assumption and Some Implications of the Simulation Argument for our Civilization* (2021), p. 3.

[146] This quote can be found in a 2017 submission on Pruss's personal blog, titled: *Are we Living in a Computer Simulation?*

[147] Bostrom (2003), p. 7.

[148] Such an undertaking is likely *impossible* given that any evidence we receive in support of our universe being non-simulated could – in theory – be simulated.

Works Cited

Aristotle. “Physics.” VIII 4-6. R. Waterfield (trans.). Oxford University Press, 2008.

Birch, Jonathan. “On the Simulation Argument and Selective Skepticism.” In *Erkenntnis*, 78, 95-107, 2013.

Bostrom, Nick. “Anthropic Bias: Observation Selection Effects in Science and Philosophy.” Routledge, New York, 2002.

Bostrom, Nick. “Do We Live in a Computer Simulation?.” In *Philosophical Quarterly*, Vol. 53, No. 211, 243-255, 2003.

Bostrom, Nick. “Why Make a Matrix? And Why You Might be in One.” In *More Matrix and Philosophy: Revolutions and Reloaded Decoded*, ed. William Irvin, 2005.

Davies, Paul C. W. “The Eerie Silence: Renewing Our Search for Alien Intelligence.” Boston: Houghton Mifflin Harcourt, 2010.

Lloyd Seth. “Programming the Universe.” Perimeter Institute, 19 Apr. 2006.

Pieri, Lorenzo. “The Simplicity Assumption and Some Implications of the Simulation Argument for Our Civilization.” OSF Preprints, 6 Apr. 2021. Web.

Brady is an honors student in philosophy at St. Thomas University (Canada), class of ’22. Currently, Brady is interested in epistemology, philosophy of science, cosmological fine-tuning, the anthropic principle, artificial intelligence, various fields of ethics, and certain historical developments within the continental tradition. Brady’s post undergraduate plans involve further education in philosophy, physics, or law.

The Neofeudal Thesis and The Frankfurt School

A Conversation with Jodi Dean, PhD

About Jodi Dean, PhD

Jodi Dean is a professor of political theory in the Political Science department of Hobart and William Smith Colleges in New York state where she has been teaching since 1993. Dean received her B.A. in History from Princeton (1980-1984) and received her PhD as well as her MA and MPhil from Columbia University (1986-1992). She taught at the University of Texas, San Antonio (1992) and held visiting research appointments at the Institute for the Human Sciences in Vienna, McGill University, and Cardiff University. Dean is a co-editor for *Theory & Event*, and has edited several volumes, including *Reformatting Politics: Information Networks and Global Civil Society*, *Empire's New Clothes: Reading Hardt and Negri*, *Cultural Studies and Political Theory*, and *Feminism and the New Democracy: Resisting the Political*. Dean has authored and edited thirteen books, the most recent being *Organize, Fight, Win: Black Communist Women's Political Writing* (2022). Dean's area of interest within political philosophy revolves around Marxism, psychoanalysis, and postmodernism. She has also recently authored several works on neofeudalism and has written extensively about Technoculture and cyberspace.

Sapere Aude: To start, where are you from and who do you think the main influences are on your breadth of work?

Dean: It's hard to say where I'm from because I moved around a lot as a kid -- but I teach at Hobart and William Smith colleges in Geneva, New York and I've lived there about 25 years. So that's I guess where I'm from. I'm in political science and I'm a political theorist, my primary – or the text and the figures who shaped my thinking the most I would say are, Lenin, Lacan, Zizek, Marx and in some ways Althusser. I'll also say, I when I first started out – Habermas. I did my dissertation on Habermas, so these are my reference points for critical theory kind of broadly.

Sapere Aude: So more recently then, you've been talking a lot about neofeudalism in your work and I think at a very basic level - what is the conceptual merit of defining this kind of state that we are in as neofeudal rather than capitalist?

Dean: Right so first - just to kind of fill out the concept a bit, my idea around neofeudalism is a response to Mackenzie Warks' question or provocation of, '*what if we're not in capitalism anymore but something worse?*'. So, I began thinking about it from this perspective of what if we're not in capitalism anymore, and that led me to think that, hey, maybe we're not in capitalism anymore. We've got, instead of the majority of economic activity being in commodity production, the majority of economic activity in services and that's not just the case in the EU, US, and in the UK but in all of the 'so-called' developed

“…it seems like more and more wealth is accumulated also through fines, fees, and rents. These are not particularly capitalist forms of wealth accumulation; they are forms of taking not making.”

countries and in a large number of the ‘so called’ developing or less developed countries. At this point, we're talking like 70-80% of the labor force working in services. So that doesn't seem particularly capitalist - it seems like more and more wealth is accumulated also through fines, fees, and rents. These are not particularly capitalist forms of wealth accumulation; they are forms of taking not making. That's an expression I get from Brent Christophers in his book on rentier capitalism which I highly recommend. So, these seem to be symptomatic of a formation that's not recognizably capitalist anymore.

So, I think about neofuedalism actually in terms of four aspects, first, the legal aspect or legal-state aspect which would be the parcellation of sovereignty. We've got lots of different mergers of the political and the economic and different forms of authority and wealth extraction throughout the social sphere where we it doesn't make sense to think in the kind of bourgeois modernity forms that these are separate. They're blurred together, that's a characteristic of feudalism. The kinds of social property relations we have now don't look a lot just like employer-worker but have dimensions of Lords and Serfs and that's like many of our relations to the platforms that kind of capture all

of our interactions and our data and metadata. Hinterlandization would be the landscape or spatiality of neofeudalism and it lets us think about the division of the kind of general social landscape into successful alpha cities and lots of desolate hinterlands. Even the division within cities between the thriving neighborhoods and the neighborhoods that have been utterly impoverished and decimated reflect this. Then finally, an affective level of generalized catastrophism and anxiety. Let's just think about the the kind of vibe or feel of neofeudalism. Those to me look really different from how anyone described bourgeois modernity and I think, thinking about our present in terms of neofeudalism lets people start to say – 'oh god, you know things are a lot worse than I thought'.

Sapere Aude: I think that that makes a lot of sense – then, in putting this thesis forward, do you think that our current neofeudal society or transitory neofeudal state is the logical extension of capitalism or say -- the height of capitalism itself? Or is it more of a returning to capitalist origins because there's no need to shroud the expropriation and exploitation of our society in something else?

Dean: Can I have it both ways? I do want to have it both ways - I want to have it both ways in that I don't want to think of it as a return because that would posit some kind of cyclical notion of history. Which, I don't think that 'going back'… the temporality doesn't sit well with me, but the way you expressed it was so good because it's not going back exactly. It's aspects of our society and our economy that

have been historically present, that are now being revealed more and more with a kind of direct presence than they held before. So, let's say forms of unwaged labor, forms of taking not making, those are present but now they're more dominant. I don't think it's a return, I think it's like a continuity from capitalist processes. It's capitalism turning itself into something else and its ongoing, right? So, it's not like my argument is not that neofeudalism has replaced capitalism - my argument is that capitalism has these neofeudalising tendencies that are now becoming dominant.

Sapere Aude: That makes perfect sense, then, for going back and reconciling or at least discussing the other fundamental parts of what modern neomarxists take to be essential for understanding the state of our social relations, how does the role of the market and ideology fit into the neofeudal thesis?

Dean: So, first, under neofeudalism we have more relations that are not mediated by the market but are mediated by direct kinds of taking. We pay fees for freaking everything, right? That's not necessarily about that fee itself, it is not the same thing as pricing, right? Weirdly, we get attached to fees for buying something. Like, if you buy tickets online for anything there's a fee for that -- which seems so strange. Or the way that when we enter into any kind of platform and they take our data and metadata, they regulate where we can go and how we can express ourselves, that's not all market relation. One last thing on the market portion of that question -- I think that things like Uber show the

"I think that things like Uber show the 'market' destroying itself. We work, they destroy markets, and they are about getting rid of the market and making it the case that in order to do X you have to do it through them."

'market' destroying itself. We work, they destroy markets, and they are about getting rid of the market and making it the case that in order to do X you have to do it through them. Or, through whatever means they provide. One of the worst versions of this is what happens to people who do things like -- maybe handymen or contractors or dog walkers or house cleaners, before they might just put their names up on a local bulletin board or rely on word of mouth. Now, we've got these digital intermediaries that come in that are the access point for a consumer looking for the service and the service provider but then they don't get to set their own terms of employment as easily as they could before. They have to give the freaking, you know, app or platform a cut. So, I think we need neofeudalism to helps us think about the way, and in fact some of the things that we've thought about U.S. markets, aren't operating actually as markets anymore.

On ideology – so, I guess it was the late 90s or early 2000s… it's hard to think about that but, people started talking a little bit about post-ideology, meaning that it's not like you could say that

there was a dominant ideology that everyone accepted or agreed with that then had to come under critique. Instead, there are multiple different, for lack of a better word, 'ideologies' -- you know, with a small '*i*'. But these ideologies, people start talking more in terms of discourses, or publics, or now identities, but to say that there's one overarching ideology doesn't seem to really fit with where we are. Like, we can recognize, "OK here are people who talk a lot about political theory… here people who more interested in religion… and here are people who, you know, talk about gardening," or whatever, but to say that everything is within one ideology doesn't capture our world.

Sapere Aude: I think that makes a lot of sense too, a good logical place to go from there might be -- how should we modify our past systemic philosophical thinking to be more reflective of our everyday activities within this neofeudal thesis? Especially given what we just talked about, because a lot of scholars that see themselves as neomarxists rely on that idea of everything as mediated by the market due to market exchange and the dominance it holds over our social relations – what do you think is the most important thing to now rewrite?

Dean: That's a smart question, I was going to give a flip answer like, "everybody should just read everything I've written and then start from there" but I don't actually think that. But, I think that what I have found kind of surprising is how interesting and appealing it is to look back at anything that was written before postmodernism and before deconstruction and to take the Marxist debates from the 70s seriously

again. It almost seems like we made this wrong turn and that neoliberalism and postmodernism was all the form of defeat and now we've got to go back to these other kinds of philosophy.

I also find really useful… I'll make a plug for the book Co-edited with Charisse Burden-Stelly, *Organize, Fight, Win: Black Communist Women's Political Writing* - like going back to this writing, the text we've collected start from 1928 and go to 1956 and this writing is amazing because this is all about the struggle and it's all about building unity. It's all about the kind of practical work of organizing against things like; white supremacy, male supremacy, imperialism, and fascism.

Sapere Aude: Why do you think that a lot of academics then resist this turn, not within just the neofeudal thesis, but resist having a dialectical conception of almost any systemic issue? A lot of academics are so committed to this tradition… not always the analytic one, but to a particular way of thinking about things?

"It's really that we've got the remnants of or are still in the wake of, anti-communism and that people need to kind of get over that and go beyond that and really appreciate that tradition."

Dean: I think it's I think it's rooted in anticommunism honestly; I think it's rooted in having been educated in either a Cold-War or post-Cold-War world that said that communism, or anything associated with communism,

was bad and defeated. “It's really that we've got the remnants of or are still in the wake of, anti-communism and that people need to kind of get over that and go beyond that and really appreciate that tradition.”

Sapere Aude: Related to that especially is, I think, the tendency to distance themselves publicly from Marxist thought in the Frankfurt school. Much of the first wave was so concerned about the optics of even saying ‘Marxism’ in their work and you can feel that same tendency in a lot of academics today. With that, how do you think that we should then conceive of this kind of transitory phase of our system within Adorno and Horkheimer's conception of the mythic and the overtly scientifically rational in Dialectic of Enlightenment?

Dean: So first, I’m going to answer this in different ways. So at the beginning of COVID, I decided to go back to *Dialectic of Enlightenment* and work through it again. As I was going through it and I felt two ways - on the one hand, like this is ridiculously hard and the other hand, I was like oh… the more I work with it, I feel the argument. I can feel it, even if I can't explain it very well. Well then, I was like, that's just an illusion! If you can't explain it very well, that’s just an illusion. So I don't really know - I feel that my

“To see all of and to read all of the problem as one of an instrumental relation to thought goes back to the problem of myth in relation to nature."

overarching sense after returning to *Dialectic of Enlightenment* is that, I think I reject their move to instrumental reason as the problem. I think that is a rejection of class struggle and a rejection of some Marxist historical materialism. To see all of and to read all of the problem as one of an instrumental relation to thought goes back to the problem of myth in relation to nature. That turn just seems like it boils down to - well, 'thinking is bad' but like that can't be what they mean. So what is this right? I mean it's like it's a trap of being stuck in a trap of thought. I just don't think that's helpful; I think they get stuck there because they abandon class struggle and then it's part of and totally becomes about; 'what does it mean to be stuck in anti-communism?'. So, I think I went off track with the question, what was the question again?

Sapere Aude: I think that that's a perfect response in line with the original question because you can see that play out in their politics in their lives, where you have Adorno calling the cops on his own students and Horkheimer being a pro-Vietnam War academic, of course that is rooted in anti-communism and the trap of being stuck in the trap of your own though and losing touch with praxis at the end of the day.

Dean: I just saw something, I didn't follow it up 'cause it was too late last night, and I shouldn't be on social media at 1:00 AM, but people were saying stuff about Horkheimer being responsible for the death of Walter Benjamin - in that Benjamin asked him for like $500 so he could get out of Germany and Horkheimer was like, 'no I don't have it'

- but he had taken something like $50,000 from the Institute for Social Research and put it in a bank. Have you seen this? I didn't follow up and who knows if it's true…

Sapere Aude: Yeah, it's hard because there's always conspiracies surrounding Benjamin's death because it's so tragic. But, I think that goes hand in hand with the conflation of their history with these complicated external politics that just totally lacked any praxis. People fundamentally don't understand what happened in their lives and you have people like Martin Jay writing a history of the Frankfurt School is not full in any way. So there's always a conspiracy about Benjamin's death that comes back to Adorno or Horkheimer but at the end of the day to blame Benjamin's death on either of them is really cruel when they were all Jewish scholars fleeing Nazi occupation and very narrowly escaping. I think all of these conspiracies attempting to place blame at all, upon anyone or anything other than the hostile takeover of all of Germany by the Nazis when he tragically took his own life, can be really reactionary and a distraction from what they were saying.

Dean: I need to follow up on this, I haven't followed all of the conspiracy theories about the Frankfurt school intently, but I see them every once in a while. Like I saw someone saying something online about connecting the Frankfurt school with the CIA I'm like, well, that's not a conspiracy everybody knows Marcuse worked for the OSS!

Sapere Aude: It's always so strange how the involvements of political philosophers on the left get scrutinized like their less-than-savory political ties

negate their work but you have people like Wittgenstein rumored to have worked for the KGB and that's seldom mentioned in any conversation about his philosophy.

Dean: Oh, that I forgot!

Sapere Aude: I did too! the last time I talked about the politics of the Frankfurt School I had this discussion about the political actions of analytic philosophers being disregarded within the rhetoric of philosophy writ large where you have people who are fundamentally anti-communist within academic philosophy that always point the finger at 'modern Marxists' and say, 'Oh well your favorite scholar worked with X' but then you look at analytic philosophers and they were doing the same or actively had ties to the Nazi party?

Dean: Well, they were all Nazi's, yeah. I guess that is not fair… yeah, no.

Sapere Aude: I mean Heidegger existed…

Dean: Well, there we go.

Sapere Aude: Well, that was a tangent but - we talked a lot about praxis today for widely different political contexts, and I think that brings in the question, should we see history as contingent or as kind of predetermined in this way that Postone articulates?

Dean: I mean I'll distance myself from the Postone part - I'll just say that, I believe this is in the 18th Brumaire where Marx writes, "men

make their own history but not under conditions of their choosing". So, both conceptions can be right. I mean I think we need to do more of this emphasizing that we're in the picture that we take. So, it's a mistake to oppose these things, I think. The other way that Zizek puts this problem is as 'subject as the gap in the structure'.

Sapere Aude: I think that makes sense and is completely in line with what you have been saying in relation to this neofeudal thesis – I guess my remaining question that's oriented more towards praxis is just, how do we conceive of ourselves within this system as we're going through this kind of transitory phase and things are presenting themselves more like servitude in this not entirely new way but very direct way do?

Dean: What I honestly think is that we've gotta stop worrying about our identities and worry about organizing to change the world.

Sapere Aude: I think that is great and reminds me a lot of what Mark Fisher wrote in Exiting the Vampire Castle.

Dean: Yeah, you know that one? I love that one. It was so powerful, I mean - that's why in my last book was called *Comrade* I dedicated it to 'MF'. I didn't spell it out 'cause, I didn't, yeah… That was a great essay, that was really like one of the first things I ever wrote a comment on, it was in response to that. I don't even know if it still exists online anymore, because it was published in *Meditations.* But yeah -- it's like, what if we stopped thinking about how do we think about ourselves and

thought about like, ‘OK what are we fighting for and how are we organizing to achieve that goal’ -- how are we organizing to fight for a better system.

Sapere Aude: Yeah and I think he was correct and you’re absolutely correct in what you’re putting forward now, I mean you saw - even in the response to that piece in academia, as soon as that became more widely discussed people were not ready to talk about identity politics and it’s inherently divisive elements when it is put before really concrete solidarity to change things. Now people are totally invested in identity politics in a really reductive way.

Dean: So it's so funny -- my very first book which came out came out in 1996 and it was called *Solidarity of Strangers: Feminism After Identity* -- I got that really wrong, right? Like - I mean, I thought it was, in the 90s at least, and it still is the case that people are talking about identity politics and the critique of it. I really thought that we were moving out of that and then instead, it kind of returned and in all sorts of different ways. I mean it seems like right now it’s useful to recognize that the right is anchoring their politics in this particular version of their own kind of mythologized white or white masculine identity. I guess that's what we talked about today too, is like, how do you escape this symbolic representation? How do you escape this fundamental reduction or like essentialization of someone’s politics? I may have mentioned already, *Organize, Fight, Win* coming out in October Co-edited by myself and Charisse Burden-Stelly, what's so great is that the

Black Communist Women are writing that are in this book, they never worried about their identities. They never worry about anyone's identity at all -- like that's not the thing, right? Instead, they might interview Black women looking for domestic work in Harlem during the depression and they talked to them about their working conditions, and they talked to them about how their you know how they're negotiating the relations with the working class white women who are trying to employ them but it never becomes about anyone's identity. The whole situation is praxis, struggle, labor, you know? Unity, that kind of thing, and I think that's useful.

Sapere Aude: Yeah, we should always just be fighting for total solidarity and find unity wherever we can, I think in that vein, the final question I have for you would be -- what unity do you see in the neofeudal thesis for praxis?

Dean: My hope is that neofeudalism as a category lets us recognize how struggles among and throughout the service sector more accurately present themselves today. Just the ecologically decimated environmental struggles, the kind of crises of social reproduction, strike struggles, the billionaires mass accumulation of wealth, struggles around technological dictation of every aspect of our lives, this is all a part of the same struggle that I think is captured by neofeudalism as a category.

Sapere Aude: I think that's a perfect close to what has been a really great conversation. Thank you for sitting down with me thank you for talking about

your current work and some of your influences, this has been very illuminating! I think this conversation really brought everything together for me, I hope it brings everything together for the people reading.