Philosophy in Practice



SPRING 2021

VOLUME 15

CALIFORNIA STATE UNIVERSITY, LOS ANGELES
DEPARTMENT OF PHILOSOPHY

Philosophy in Practice

VOLUME 15 – SPRING 2021

© Copyright 2021 by CSULA Philosophy Department. All rights reserved. Except for brief quotations in a review as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a data base or retrieval system, without the prior written permission of the publisher. Individual copyright reverts to the authors upon further publication of their articles.

Contents

Acknowledgements iv

Faculty v

Professor Spotlight: Henry Mendell vii

Articles:

The Viability of Physicalist Emergentism
Stephen Boynton 1

Persons Evereywhere Everywhen and One of Each: Motivating Toward Observer Presentism Panpsychism
Daniel Castro. 23

The Right to Choose a Deaf Child
Alexandra Meyering 30

Unity of Virtue and Wisdom in the *Protagoras*
John Hurley 53

Human Rights, Universality, and Moral Desert
David Randall 70

A Question of Control
Heather Norwood 94

Hutto and Myin Fail the Hard Problem of Content
David Dixon. 109

The Idols of the Mind in Modern American Political Economy
Sabrina Pirzadaa. 135

Doing Our Duty: The Moral Impermissibility of Suicide
Hudson Olander 147

The Film-as-Philosophy Debate
Marcel Giwargis 163

Can Happy Hard Determinists Stay Happy?
James Savage. 175

Contributors 194

Philosophy Program Information 196

Acknowledgements

Philosophy in Practice is produced by students at California State University, Los Angeles. The editorial staff wishes to thank all the authors for submitting and refereeing articles, the faculty of the philosophy department for agreeing to supervise individual contributions, and the University for its generous financial support.

Editors:

Stephen Boynton
John Hurley
Heather Norwood
Sabrina Pirzada

Faculty Advisor:

Dr. Michael K. Shim

California State University, Los Angeles
Philosophy Faculty

David Pitt (2003–), Chair, Ph.D. City University of New York. Philosophy of Mind, Philosophy of Language, Metaphysics

Mohammed Abed (2008–), Ph.D. University of Wisconsin, Madison. Ethics, Social and Political Philosophy, Philosophies of Violence, Genocide and Terrorism

Bruce Atta (1996–), M.A. California State University, Los Angeles. Metaphysics, Epistemology, Ethics, Social and Political Philosophy

Mark Balaguer (1992–), Ph.D. City University of New York. Philosophy of Mathematics, Metaphysics, Meta-ethics, Philosophy of Language, Logic

Talia Bettcher (2000–), Chair, Ph.D. University of California, Los Angeles. History of Modern Philosophy, Philosophy of Self, Gender and Sexuality

Jay Conway (2005–), Ph.D. University of California, Riverside. 19th-20th Century Philosophy, Modern Philosophy, Aesthetics, Social and Political Philosophy

Richard Dean (2009–), Ph.D. University of North Carolina, Chapel Hill. Ethics, Kant's Moral Philosophy, Applied Ethics

Foad Dizadji-Bahmani (2013–), Ph.D. London School of Economics, United Kingdom. Philosophy of Science, Philosophy of Physics, Philosophy of Probability

Ronald Houts (1983–), Ph.D. University of California, Los Angeles. Metaphysics, Epistemology, Logic

Keith Kaiser (2001–), Ph.D. University of California, Los Angeles. Metaphysics, Philosophy of Mind

Steven R. Levy (2005–), Ph.D. University of California, Los Angeles. Epistemology, Metaphysics, Logic, Applied Ethics

Henry R. Mendell (1983–), Ph.D. Stanford University. Ancient Philosophy, History of Ancient Mathematics and Science, Philosophy of Science, Metaphysics

Sheila Price (1964-), M.A. University of California, Los Angeles. Recent Philosophy, Comparative Religions, Medical Ethics, Environmental Ethics

Michael K. Shim (2007-), Ph.D. State University of New York, Stony Brook. 20th Century Continental Philosophy, Phenomenology, Husserl, Modern Philosophy, Philosophy of Mind

Emeritus Professors

Thomas Annese (1961–1992), Epistemology, Modern Philosophy

Sharon Bishop (1967–2004), Ethics, Political Philosophy, Philosophical Psychology, Feminist Ethics

Donald Burrill (1962–1992), Ethics, Philosophy of Law, American Philosophy

Ann Garry (1969–2011), Feminist Philosophy, Philosophical Methodology, Epistemology, Applied Ethics, Wittgenstein, Philosophy of Law

Ricardo J. Gómez (1983–2011), Philosophy of Science and Technology, Philosophy of Mathematics, Kant, Latin American Philosophy

Joseph Prabhu (1978–2017), Philosophy of Religion, 19th and 20th Century German Philosophy, Moral and Social Philosophy, Indian and Comparative Philosophy

Kayley Vernallis (1993–2019), Moral Psychology, 19th and 20th Century Continental Philosophy, Feminist Philosophy, Ethics, Aesthetics, Gender and Sexuality

George Vick (1967–1997), Metaphysics, Phenomenology, Existentialism, Philosophy of Religion, Medieval Philosophy

Professor Spotlight: Henry Mendell



Professor Henry Mendell knew at age 13 that he would be a philosopher. Mendell developed an interest in analytic philosophy as a teenager, and went so far as to attend an academic philosophy conference as a senior in high school. His lifelong interest in mathematics and his study of Attic Greek before college led him to specialize in ancient philosophy. Much of his work is on Aristotle and ancient mathematics.

In addition to his B.A. in Classics and Philosophy from Cornell, Mendell earned a second B.A. in Philosophy at Cambridge University. He attended Cambridge "by pure accident"—he'd applied for a B. Phil. (master's level equivalent) program at Oxford and a second B.A. program at Cambridge, but his Oxford application was lost due to a postal strike. Cambridge turned out to be the right choice though. Mendell already knew he wanted to focus on Aristotle, and he was able to work with leading Aristotle specialist G.E.L. Owen at Cambridge. He joined the philosophy department at Cal State LA in 1983, obtaining a tenure-track position several years before finishing his Ph.D. thesis at Stanford.

Dr. Mendell describes himself as a "hyper-historian" in his view of ancient work. Asked which area of ancient philosophy is most relevant to the modern world, he first resists the question on the grounds that it can lead to misreading of ancient texts. Ancient work like Aristotle's *Physics* is valuable for the insight it provides into ancient thought, he suggests, rather than for substantial understanding of physics.

"Part of the purpose of studying the history of philosophy

is to learn what it is for things that are familiar to actually be very unfamiliar." Learning ancient philosophy is "a process of defamiliarization rather than familiarization." It is easy to read modern thinking back into ancient texts, but "there are areas of ancient thought that are hopelessly not modern." Students reading Aristotle for the first time will find much that is familiar, because Aristotle "is part of the modern world." Yet, to understand Aristotle "as it is," they must put the modern context aside. Failure to do this results in misreading of ancient texts. For instance, "Nietzsche picks up pre-Socratic philosophers [in an ahistorical way] and decides that there is a Schopenhauerian way of looking at them. Other writers have interpreted the work of the geometer Apollonius in terms of algebra, which was unknown in ancient times. They may succeed in rewriting his insights in a formally equivalent way, but they miss the value of understanding Apollonius's way of thinking." Dr. Mendell believes we should not read ancient work only to understand how it led to modern thought, but on its own terms; ancient texts can reveal "what the possibilities of human thought really are" only if we understand them in their own context. "We don't want to remove the familiarity entirely, but making it unfamiliar is a useful exercise in seeing how human thought looks when it's contextualized, while also realizing what is non-contextual about it." The tasks of reading an ancient text is not just to understand its relation to modern thinking, but also to see "both abstractly and in the details how it is a different way of thinking… that is important and valuable. And that is the hardest thing to do."

"When you read a text with an argument in it," Mendell says, "there are always gaps in the argument. Where there is an important concept being used, it often looks like there is an amphiboly [a semantically ambiguous phrase]; in this place [an ancient writer] is using a concept in this way, in another place he's using it in another way." Yet, what looks like equivocation to us in fact reflects our understanding of a distinction that did not

exist for an ancient writer. Once the distinction is "in the air, no one knows what it is to think without that distinction. Then people sit down to write about history, and they don't know how to write about those writers without their modern conceptualization." For example, at some point the Presocratics began discussing the relationship between appearance and reality. The sophists Protagoras and Antiphon may have rejected the distinction, claiming that what we perceive is reality itself. Yet, once the distinction was part of the common language of thought, their position could not be to think without the distinction; it could only be "thinking with the distinction, but rejecting it." The question for us in trying to understand texts on their own terms is "what is it to think without the distinction? Is that something possible?"

It is in the field of ethics that ancient texts offer us substantial ideas that are most relevant today. "Aristotle's ethics are still very much alive," he says, and Epicureanism and Stoicism provide approaches that remain worthy of consideration. Ancient ethics provides "ways to remove oneself from the chit-chat of one's own world... When one reads ancient ethics one is put in a different world and so it gives you a way to see the abstract and separate it out from the clatter of ordinary common discourse about how to lead one's life." We read Stoicism or Aristotle and are repelled by arguments that we should accept slavery or that women should be subordinate to men, but it is important to focus on the principles, not the specific prescriptions. "You can see the forest outside of those trees, and that is part of what studying ancient texts does."

Despite his interest in ancient thought, Mendell is not a technophobe. During graduate school he completed a substantial software development project, creating a text editor that could handle footnotes and Greek text. Yet he is concerned about the loss of intellectual breadth for scholars as electronic text replaces paper books and, particularly, brick-and-mortar libraries. He has found great enjoyment in books encountered while browsing library shelves, including a "wonderful" Ph.D. dissertation on

obscenity in German translations of Shakespeare and an "insane" one arguing that religious minorities in Pakistan would be better off under Islamic religious law.

Dr. Mendell currently teaches part-time as part of the Faculty Early Retirement Program. He will retire from teaching at Cal State LA at the end of the 2021-2022 academic year. He'll be teaching an undergraduate course on ancient philosophy in the fall and Aristotle in spring 2022. Ancient philosophy enthusiasts will not want to miss the grand finale of his Cal State LA career.

The Viability of Physicalist Emergentism

Stephen Boynton

Introduction

How can the mind be understood in physicalism without reducing it to brain matter or eliminating it from the picture entirely? Contemporary proponents of so-called "physicalist emergentism" think they have the solution: a theory that takes physicalist ontological monism seriously, avoids the pitfalls of other forms of nonreductive physicalism, and allows for a not only real but causally efficacious mind. If the theory is to be viable, however, it must have an answer to Kim's (1999) causal exclusion argument. Zhong (2020) thinks we can reject causal closure as narrowly construed by Kim in favor of a broader definition, allowing for efficacious emergent properties while still staying grounded in physicalism. Megill (2012) trades emergent properties for emergent entities that are just as physical as the brains from which they emerge.

If either of these solutions were viable, the results would be tremendous. But neither solution is: physicalist emergentism, at least as a theory of the nature of mind, collapses in on itself. Zhong's account succumbs to other causal problems and while he can possibly hang on to physicalism, what surfaces is not emergentism in the sense he has in mind. Megill's emergent entities can't actually get the grounding they need to be anything more than "physical" in name only, and so he might be able to maintain his emergentist commitments but only by abandoning physicalism. There's no way to thread the needle such that one can be an emergentist about the mind while also being a physicalist.

In this paper, I will conclusively show why the position is

untenable. First, in Section I, I'll provide some definitions, so it's clear what's required by emergentism generally and physicalist emergentism more specifically. Then in Section II, I will rehearse Kim's formulation of emergentism and sketch his causal exclusion argument against it. Having done so, in Section III, I'll consider Zhong's solution: the denial of "narrow" causal closure. In Section IV, I'll discuss what I'll refer to as "Lewtas's Dilemma:" a set of new causal arguments against emergentism, which show that while Zhong may be able to hold onto physicalism, he can't make sense of emergent causal powers. We'll then turn to Megill's denial of property emergence in favor of emergent entities in Section V. Finally, in Section VI, I'll discuss the grounding problem that afflicts Megill's theory, show that there's no available solution to him, and that without grounding, his emergent entities can't be accommodated to physicalist ontology.

Section I: What We Talk About When We Talk About Emergence

Before we can consider Zhong and Megill's particular formulations, we should get clear on what's meant by emergentism generally and then physicalist emergentism more specifically. What emergentism amounts to is frequently unclear. As a case in point, many recent papers present taxonomies in attempt to delineate the different forms that emergentism can take (See Welshon, 2002; Sartenar, 2015; Paoletti, 2017). Chalmers (2006) broadly divides emergentism into strong and weak forms. *Weak emergentism* is the claim that there are emergent properties of complex systems that are "unexpected"—that is, they're not practically deducible from information about the lower level or "basal conditions." As Chalmers notes, much of the talk of emergence in philosophy of science has this concept in mind (Chalmers, 2006).[1] Our interest here is not in weak emergence but strong or *ontological emergence*.

While this will be sharpened below, ontological emer-

gentism can roughly be understood as the conjunction of two claims: (1) certain organizations of lower level entities give rise to *novel* "emergents"[2] that have causal powers not possessed by the basal conditions from which they emerge; and (2) these emergents are ontologically discontinuous with, and irreducible to, the basal conditions from which they emerge. Physicalist emergentism, specifically, takes this idea and tries to accommodate it to physicalist ontology. What this amounts to, for a theory of mind, is the above two claims conjoined with (3) the mind (as emergent) is as physical as the basal conditions in the brain from which it emerges.[3]

Why is this position interesting? It promises something tremendous. Sartenar (2015) describes what he calls the three promises of physicalist emergentism: the ontological promise, the epistemic promise, and the theoretical promise. The ontological promise is that "emergence would be an efficient tool to conciliate some form of scientifically respectable monism with the safeguard of some important bits of reality against reduction and elimination" (Sartenar 2015, p. 81). The epistemic promise provides an account for and buttresses the autonomy of the special sciences. Finally, the theoretical promise is that physicalist emergentism could provide an avenue for solving major philosophical problems, most relevantly (for our purposes) the mind-body problem. Neither dualism nor standard formulations of physicalism can offer all three, so this alone is reason enough for seriously engaging with the position.

Section II: The Causal Exclusion Argument

As stated at the outset, if emergentism is to be viable, it must answer Kim's causal exclusion challenge. Before sketching the argument—which applies to any form of nonreductive physicalism—it's necessary first to understand how Kim formulates the

emergentist position that he applies his argument to, as well as some important background principles under which he operates.

Firstly, any physical object has what Kim calls a "total microstructural property": (i) the particle constituents of the system, (ii) the properties of those particles, and (iii) the relations between those particles. The total microstructural properties are the basal conditions of the system. Emergentism thus assumes "a layered view of reality" wherein there are different levels of objects: e.g. the brain (the total microstructural property of which forms the basal conditions) is at a lower level than the mind (the emergent). Kim takes it that emergentists accept mereological supervenience: "systems with an identical total microstructural property have all other properties in common" (Kim 1989, p. 6). Emergents must then be distinguished from "resultants," which are predictable from the basal conditions (at least in principle), whereas emergents are not. Though emergents are not theoretically predictable from basal conditions, they are inductively predictable, allowing for the formulation of emergence laws. Since this implies only nomological necessity between the emergent and its base, emergence laws state "a brute correlation between M [microstructural property] and E [emergent]" (ibid, p. 9).

To understand why emergents aren't reducible to their basal conditions, we need a model of reduction. Kim here offers the functional interpretation of properties: a higher-level property can be reduced to its base by identifying the basal properties that do the causal work of the higher-level property. The property, or properties, in the base that do the causal work are the "realizers" of the higher-order property. This leads to the critical *causal inheritance principle:*

> *If a functional property E is instantiated on a given occasion in virtue of one of its realizers, Q, being instantiated, then the causal powers of this instance of E are identical with the causal powers of this instance of Q (ibid, p. 16).*

Kim operates under "Alexander's dictum:" that to be is to have causal powers—thus two properties (e.g. the mental properties of the mind and the corresponding physical properties of the brain) with identical causal powers are the same property. So, if we can "functionalize" some higher-order property—that is, construe the property "as a property defined by its causal/nomic relations to… properties in the reduction base" (ibid, p. 10)—then, by the causal inheritance principle, the higher-order property reduces to the lower-order property.[4] However, emergents, as properties with novel causal powers, cannot be functionally reduced. An emergent property thus must be able to causally effect other emergent properties, its basal conditions, or other properties at the level of its basal conditions. But emergents are also, of course, *determined* by their basal conditions. The conjunction of these two principles—downward causation and upward determination—creates a problem: how can something simultaneously be determined by a thing and effect substantive change in it? As Kim puts it, "if causation or determination is transitive, doesn't this ultimately imply a kind of self-causation, or self-determination—an apparent absurdity?" (ibid, p. 28).

The solution is to drop the requirement of simultaneity. Emergentists can avoid causal circularity and emergents can still be robust causally efficacious properties if downward causation is understood diachronically. Let P stand for the total microstructural property of the brain at time t_1, M the emergent mental property determined by P, and P* the microstructural property of the brain at time t_2. This is how the diachronic case is supposed to work: at t1 P obtains which causes M to emerge (upward determination), then the emergent M exerts downward causal force (downward causation) on P such that P becomes P* at t2. Note that the physicalist emergentist is specifically committed to this account: it must be that M causes P to change to P* rather than M causing some other mental state, or otherwise we'd be in the territory of Davidsonian Anomalous Monism (or worse, dualism). As Kim

acknowledges, on its face this account is perfectly plausible. But it is here that causal exclusion rears its ugly head.

Any form of nonreductive physicalism that's not a species of epiphenomenalism (which emergentism of the form we're discussing certainly isn't) assumes *Psychophysical Causation:* "an event, *in virtue of its mental property*, causes another event to have a certain physical property" (Kim 1989, p. 43). But for an ontological theory to rightly be called physicalist, it seems it must endorse causal closure of the physical (*Closure*): every physical event has a sufficient physical cause. So P*, from above, must have a sufficient physical cause. But M, by assumption, is a mental property, so M can't be the sufficient cause of P*. P, however, is a physical property, so P must be the cause of P*. Well then, how does M figure in here? We might say that P and M are each partial causes of P*, but this would violate causal closure, as it would mean part of P*'s sufficient cause is non-physical. Could we say they are both fully sufficient causes of P*? In this case, P* would be *overdetermined*. This would be problematic for at least two reasons. First, the solution is just implausible. It implies that overdetermination is rampant and systematic: in every single case where we want to say a mental event causes a physical one, we have a case of an overdetermined effect—i.e., any given physical state like P* has two sufficient causes. Secondly, this potentially puts us in violation of Closure again: if P* is overdetermined by P and M, then in the counterfactual where P is absent but M is present, M would be enough on its own to cause P*, meaning P* would have no sufficient physical cause. Overdetermination is thus no solution. Well then, how about suggesting a causal chain where P causes M which in turn causes P*? In this case we may be able to accommodate the requirement of Closure (P remains the sufficient physical cause of P* with M being a necessary intermediary), but the relation between P and M is a supervenience relation, so it doesn't make sense to talk about P causing M as if it were a separate entity. Since we can't reject Closure if we're

to be physicalists, we must reject Psychophysical Causation. The result is that the emergent is rendered totally causally inert. But then the core of emergentism is undermined and we must abandon the position.

Section III: Physicalism Doesn't Give Us Closure

Kim's argument is airtight. If you accept both his key premises then physicalist emergentism so formulated is done for. The only premise we could possibly reject is Closure. Though it seems like this premise is at the core of physicalism: every physical event in our ontology better have a physical cause. So, at first blush, giving up Closure sounds like it means giving up physicalism entirely.

But does physicalism actually require Closure as defined in Kim? Zhong (2020) argues that Kim's principle of Closure is too restrictive and that we can reformulate Closure such that it allows for things like causes in virtue of mental properties without contradicting the core tenets of physicalism. If this is right, then the causal exclusion argument won't go through, and there's thus an account of physicalism wherein we can make sense of causally efficacious emergents.

Let's first examine what purportedly motivates Closure. Physicalism, of course, takes physics as an at least approximately correct view of reality and it's standardly thought that the law of conservation of energy provides empirical support for Closure. To wit: "…energy and momentum [are] conserved within the physical domain. The occurrence of any physical event at t_1 will involve change in the amount of energy and momentum, and this change is sufficiently attributed to change in energy and momentum that accompanies earlier physical events at time t_0" (Zhong 2020, p. 37). Obviously, this alone doesn't give us Closure, as Closure is specifically about the types of causes allowed in physicalism and the law of conservation of energy says nothing about causation per

se. To get Closure, one must supplement the law with an account of causation, such as the *conserved quantity theory of causation*, which understands causal relations just as "relation[s] between events that transmit conserved quantities such as energy and momentum" (ibid, p. 37). If this is the correct account of causation, then Closure follows naturally: under this theory causation just reduces to energy transfer, so of course all events have sufficient physical causes in virtue of physical properties. As Zhong notes, however, this is a problematic view of causation. To see why, imagine a possible world where the law of conservation does not hold: sometimes energy is lost or gained without being transferred. This would be a very different world from ours but it's not logically incoherent and thus not metaphysically impossible. But if causation is just energy transfer, then we're committed to saying there's no such thing as causation in such a world, which seems dubious at best. I take this as reason enough to think that causation can't be reduced to energy transfer. But if the law of conservation of energy isn't enough by itself to explain causation, then it can't provide an inference to Closure.

Further, the law of conservation of energy isn't in contradiction with the falsity of Closure. To see this, consider the case where I intend to get a drink of water (M1) at t_1 realized by a physical brain state (P1) and at t_2 I am taking a drink (P2). As I'm aware of the importance of proper hydration, but also a lover of efficiency and symmetry, I have two water bottles on my desk, each equally full and equidistant from each of my hands. To realize P2, I could reach out to the left bottle with my left hand or to the right bottle with my right hand. Assume in both cases the same amount of energy would be transferred. My intention (M1) is to drink some water: pure and simple. But depending on whether I'm about to reach out my left or right hand to grab either the left or right bottle, there will be distinctly different activity in my motor cortex. This doesn't change my intention—in fact, I'd wager that with this sort of set-up, exactly which bottle I go for is entirely an

unconscious decision. So, M1 is then realization-insensitive: it's the same despite the particular patterns of activity in my motor cortex.[5] It seems then that we should say M1 is really the cause of P2 and not P1: M1 is of course dependent upon *some* physical state, but the point here is that as long as I have the intention, I will be taking a drink of water, and that there can be substantive differences in the physical states that realize M1 which don't substantively change M1. Note that nothing I've said is in conflict with the law of conservation, and yet this seems like a case in which explicating the cause only in terms of physical properties seems, at least, insufficient. To make sense of this, Zhong introduces a distinction between the cause *per se* and the *grounding* of that cause. Since Zhong's account is physicalistic, any given mental property must be physically grounded: in this case, that means that M1 is essentially dependent on P1. But since mental properties, as emergents, aren't reducible to their basal conditions, we can separate the causal power of a mental property from its physical grounding. So, in the above example, M1 is the cause of P2 while P1, as the physical grounding of M1, is the grounding of the cause: we need P1 to obtain to get M1, but M1 is doing the essential causal work. But even if we say M1 is grounded by P1, labeling M1 as the cause of P2 puts us in violation of Closure. This is because Closure, as construed in Kim, demands not just that every physical event have a physical cause but that every physical event is caused by another physical event *in virtue of* its physical properties. But why should we accept this? As we've seen, the law of conservation isn't enough to explain causation on its own, and it isn't in conflict with the falsity of Closure. Further, as we saw, since M1 is realization-insensitive between different physical states, we have compelling reasons to think M1 is the cause rather than P1. Here is how Zhong sums all this up: "Even though mental-physical causation (such as the M1-P2 causal relation) is physically grounded, the fundamental physical fact that grounds the causal relation is itself not a *causal* fact. That is, the

cause of P2 is M1, rather than P1 or any other physical property. Grounding doesn't entail reduction." Thus it doesn't seem that we need to be wedded to Closure, as formulated in Kim, in order to retain physicalism.

Even so, we certainly need some way to keep disembodied Cartesian ghosts out of our ontology if we want to have any claim to our theory's being physicalist: i.e., some sort of causal closure of the physical *is* necessary. So rather than dispensing with Closure whole cloth, Zhong introduces a modified version of Closure, which he calls *Closure**. The principle is the same except we trade "physical" for "physical*:" a broad sense of physical that can accommodate mental properties that are physically grounded. Is this satisfying? Personally, I think it's right to be suspicious of this move: Zhong seems to be trying to define the problem away and doesn't really do enough to explain the difference between physical in the narrow sense (what I'll refer to as "physical-" for clarity) and mental properties. And as I'll show in the following section, when we consider the implications of how to flesh out this distinction, Zhong runs into a whole new causal conundrum.

Section IV: Lewtas's Dilemma

Qualia, the individual subjective qualities of experiences, are canonical mental properties; Zhong wants to say that they have causal powers that differ from the physical base they are grounded in. So, let's consider what his account would say about a pain quale. I have been typing for hours and my right wrist is beginning to bother me. Nocioreceptive neurons in my wrist are firing, relaying signals to neurons in my spinal cord, which are then sending a signal up to my brain, causing some group of neurons to fire in my anterior cingulate cortex. All of this is happening in virtue of physical properties: the neurons are firing because the displacement of charged sodium and potassium ions is changing the membrane potential of the neurons, causing action potentials.

Now, somehow, from this process, a pain quale emerges (M1), which is an emergent mental property and thus is ontologically distinct from those neurons firing (P1). According to Zhong, it's that pain quale which does the real causal work of making me get up from my desk for a moment to stretch (P2),[6] though something's clearly missing here. How does M1 exert its causal force on my physical brain to result in P2? We're of course barred from saying it's from fundamental physical interactions: M1 is a property that emerges from and is grounded by physical properties, but is ontologically discontinuous with and irreducible to the physical properties that form its basal conditions. So how does M1 act on my physical neurons? This problem forms the core of what I call "Lewtas's Dilemma," after Lewtas's (2017) argument that, given a picture like the above, there's no plausible way to make sense of the emergent quale having any causal power.

Lewtas considers two different models of causation that the emergentist might appeal to in order to explain how emergent mental properties can cause changes in physical properties at the level of the emergent's basal conditions: "active causation" and "passive causation" (Lewtas 2017, p. 477). I'll briefly say a word about passive causation but don't think it's worth spending much time on, as it doesn't apply to Zhong's account. As Lewtas explains it, the idea of passive causation is that "the conscious property sits there, with the intrinsic nature it has, doing nothing, not exerting any force… [and] the relevant physical entity actively responds to it… and change[s] its own state accordingly" (ibid, p. 480). This may not sound much like what most people mean when they talk about causation, but it's not incoherent. Coherent or not, it's certainly not what Zhong has in mind. Consider the pain quale example: a series of physical events leads to activity in my anterior cingulate (P1), which causes a pain quale (M1) to emerge, and then *that* downwardly causes things to happen in my brain that make me get up and stretch (P2). Maybe we need to add an intention to get up and stretch (M2) that comes between M1 and P2—

or, to be precise, we'd have to say that M1 then causes a different physical state, that then causes M2 to emerge—but it's certainly not that my brain is "reading" my conscious intention and then acting accordingly. No, in Zhong's account, my conscious intention, as a mental property, causes me to get up and stretch. Other emergentists may find the idea of passive causation attractive, but it's not relevant for us and that's all that needs to be said about it.[7]

What Zhong has in mind is what Lewtas dubs "active causation." Active causation is what we standardly think of when we think about causation: it's the sort of causation operative between objects in virtue of their dispositional properties, or what Lewtas calls their "causal/functional" properties (ibid, p. 478). The cue ball is hit by a pool cue, it collides with one of the billiard balls, and that *causes* it to move based on the force carried by the cue ball, the mass of the two balls, the friction of the felt, etc. What imbues these objects with their causal/functional properties? The interaction of fundamental physical forces, of course: the mass of the balls results from the mass of the particles that make up the balls. Or, more to the point: causal/functional properties result from physical properties. But we know Zhong considers mental properties as something distinct from physical properties. Thus mental properties cannot be causal/functional; if we were to treat mental properties as causal/functional, then they would be explicable in terms of and reducible to fundamental physical interactions. So the emergent mental property itself has to be non-causal/non-functional. But maybe we can say that the emergent, somehow, exerts its causal power via familiar causal/functional mechanisms. Concerns of how this would even work aside, as Lewtas points out, "nothing essentially non-causal/non-functional can equal anything essentially causal/functional" (ibid, p. 479). So, accepting that the emergent conscious property is non-causal/non-functional, while the emergent causal power is causal/functional, means the emergent mental property must be *distinct* from the emergent causal power. "[And] if the power stands over and

above the conscious essence, existing in addition to it, then nothing stronger than nomological bundling keeps the two from pulling apart" (ibid, p. 479). The connection between emergent mental property and emergent causal power is thus reduced to a mere correlation. All of this together means that the property itself—the quale—is playing no causal role: the emergent property is rendered epiphenomenal. But this means that Zhong's emergents can't downwardly cause anything. Zhong may get around Kim's causal exclusion argument by denying Closure, but Lewtas's Dilemma shows that he ends up in the same place anyway.

Section V: Divesting From Insolvent Properties

Starting with the assumption that mental properties are distinct from physical properties, Kim's argument can be distilled into four parts:

(1) Psychophysical Causation: mental events sometimes cause physical events

(2) Closure: every physical event has a sufficient physical cause

(3) Non-overdetermination: physical events are not over-determined

(4) Exclusion: no mental event can be the cause of a physical event

Obviously (1) and (4) are inconsistent. Since we've seen that rejecting (2) runs us into Lewtas's Dilemma, it seems again that we're better off rejecting (1). But there's an unstated background premise in the argument that comes from a number of assumptions in Kim's formulation:

(5) Property Dualism: mental properties are at a *higher-level* than physical properties.

How does this premise operate in Kim's argument? Firstly, recall

that Kim took it that emergentists hold to a layered view of reality: this isn't wrong so much as it needn't commit us to property dualism. There's a sense in which cells are at a higher level than molecules, but no one would say that cells are a *property* of an arrangement of molecules; cells just *are* a particular ordered arrangement of molecules. The "levels," in this sense, don't have any ontological import; they're just explanatorily useful. In nonreductive physicalism, however, the mental is not just the physical, despite their obviously being related. The standard way of understanding the relation is via mental-physical supervenience: the mental supervenes on the physical and this gives way to levels that do carry ontological weight. Kim assumes that emergentism is committed to supervenience as well, because this is commonly how the mind is *realized* in the brain in nonreductive physicalist accounts of mind. But the ontological emergentist isn't strictly committed to realization. Megill (2012) takes it that the ontological emergentist must hew to three key tenets: the emergent "(i) stands in a lawful relation to its emergence base, (ii) has novel causal powers not had by its base, and (iii) is irreducible to its base" (Megill 2012, p. 599). Consider this in relation to Kim's causal inheritance principle, which states that if a functional property is realized in virtue of its basal conditions, then the causal power of the property is identical with those of its base. The implication, of course, is that if a mental property is realized in the brain, then the causal power of that mental property is identical with its neural base—something that's inconsistent with (ii). So, Megill suggests, to make sense of ontological emergence in a physicalist framework, we must reject the claim that the mind is realized in the brain, thus jettisoning both strong supervenience and property dualism. Pointedly, this distinguishes Megill's emergentism from other forms of nonreductive physicalism, which are structured around the realization requirement: functionalism, for example, is a form of nonreductive physicalism just because of its commitment to multiple realizability. In rejecting realization,

Megill's emergentism also rejects multiple realizability, or any sort of token-identity claim. But if we reject realization, and thus property dualism, then how can we make sense of the idea of an irreducible mind at all? Doesn't this just mean accepting reductive physicalism?

Megill agrees that minds are physical in the regular sense but denies that they reduce to the brain. It's here that there's a disanalogy with the sense of levels between say, cells and molecules. The cellular "level" reduces to the molecular, and the molecular to the fundamentally physical. Emergent minds, on the other hand, while as physical as brains or cells or rocks, do not reduce to the brains from which they emerge, as required by condition (iii). This isn't *prima facie* incoherent, but we might reasonably wonder if there are any physical things that are like this. One example might be entangled pairs of subatomic particles: some philosophers of quantum mechanics think entanglement only makes sense in terms of emergence, with the pair being an emergent that doesn't reduce to the properties of either of the constituent particles in the pair (see, for example, Lewis (2016)). And whatever an emergent pair is, it's certainly physical. But what about something that could work as a candidate for the mind? Megill suggests the EM field of the brain might be just the thing. The brain's EM field is a well-defined physical phenomenon—it's what makes EEG possible—and it's emergent in the sense that it results with law-like necessity from the synchronized firing of neurons—satisfying condition (i)—but is distinct from the firing of the neurons themselves:[8] that is, it is something over and above them (quite literally), satisfying condition (iii). There is also evidence to suggest that it might alter neuronal firing, thus demonstrating downward causation and so meeting condition (ii).[9] We might object, however, that while this shows the EM field is a physical emergent in the right sense, nothing shows that it's the mind. However, the idea has some empirical support—particularly from McFadden (2002), who argues the EM field is a partic-

ularly good candidate for a correlate of consciousness, offering up both empirical evidence in support of the idea and a number of testable hypotheses to determine if alterations in the EM field reliably produce alterations in consciousness. And anyway, Megill needn't commit himself to the EM theory of consciousness—it's a useful proof of concept, which is really all he needs.

We might still think that this isn't quite enough to save emergent entities from causal exclusion, however. Recall the picture Kim sketched: the presence of basal conditions P causes M to emerge, and then M causes P*.[10] But what causal role can M really be said to be playing here? What's at issue is the transitivity of nomological sufficiency: "If P is nomologically sufficient for M, and M is nomologically sufficient for P*, then P will be nomologically sufficient for P*… and so [P] will count as its cause, thereby making M epiphenomenal." (Megill 2012, pp. 605) But, as Megill shows, this isn't really much of a problem. Note first that the emergentist needn't—and shouldn't—claim that M alone is sufficient to cause P*; instead, he should claim that it's both M and P that are necessary to cause P*. To see why this is substantive, imagine the counterfactual where P occurs but M doesn't. In this case, P* is not caused because without the causal powers of M, P* will not come about. Put succinctly: nomological sufficiency needn't imply causal exclusivity. The emergentist can answer Kim by saying that both P and M are necessary to cause P*, that together they make up the sufficient cause, and in the counterfactual where either one is absent, P* will not come about. Since M is physical just like P, there are no worries about violating Closure. And Lewtas's Dilemma doesn't apply: the "mental" is able to flex its causal muscle in the same way anything else is. It seems then that Megill's account avoids the causal quagmire that property emergence theories (like Zhong's) fall into. But how is it that something in no way realized in the brain can have the right kind of connection to the brain that the mind clearly has? In the final section I turn to this "grounding problem."

Section VI: Tough Realizations

So far we have discussed Megill's theory in relation to the mind, or consciousness, considered as a whole. But his account must of course also explain how qualia fit in as properties of this emergent whole. So how about our pain quale then? Well, it's just a physical property of a physical entity: the emergent mind. We may think it's odd to consider qualia physical properties, given the unique qualities of qualia, but there's no contradiction here. And uniqueness or strangeness alone isn't a condition for thinking something must be non-physical.[11] Subatomic particles behave in ways that are markedly distinct from the macroscopic physical things we regularly interact with, but no one would deny that electrons are physical entities or that spin is a physical property.

There's a deeper problem, however: how is it that any quale—as a property of my emergent mind—can have any connection with what's going on in my brain? I am now having a visual perception of my keys on my desk and am thinking of stepping outside of my apartment to get some air. There's the conscious percept, which Megill would locate in the non-realized mind, but there's also a certain pattern of activity in my visual cortex. Here Megill actually sees an advantage for his theory: he notes that qualia are particularly problematic on theories that take realization seriously because of the bifurcation of visual processing into the ventral and dorsal streams. As is well-known, after basic feature processing in V1, visual information is fed into the ventral stream for the construction of coherent object percepts, and into the dorsal stream for behavioral outputs related to these objects. The percept of my keys is a result of ventral stream activity; my reaching out to grab them requires that same raw sensory data go through the dorsal stream. The problem is that the streams are segregated, so it seems as if the coherent object percepts put together in the ventral stream aren't actually related to behavior which, if we want to say qualia are behaviorally relevant, is indeed a problem. Megill notes

that localizing consciousness—and thus qualia—in an emergent entity that is not realized in brain tissue solves this issue: qualia are behaviorally relevant because they are properties of emergent consciousness. It doesn't matter that their basal conditions are in the ventral stream, it all gets put together in the emergent whole that exerts downward causal force on the brain.

But if consciousness isn't realized in the brain, in what sense can activity in the ventral streams be tied to specific properties of the emergent at all? This is how Schroeder (2015) describes the issue, which he refers to as the "grounding problem":

> The problem is that… we do not need lower level physical properties at all. There seems to be no reason to even talk about the ventral or dorsal systems in this case. Independent from either system, the higher-level property just occurs somehow; possibly from somewhere else… Because it is non-realized, locating it in some region of the brain is speculative, or correlative at best. Even if it does seem to always occur in the presence of a particular physical realizer, it need not be the case… Without the realization requirement, emergent properties have no accountability to the causal structure of the physical world. (Schroeder, 2015, pp. 490)

And if something is unaccountable to the rest of the physical world, it's hard to see how it can actually be physical in any meaningful sense.

Schroeder is interested in saving Megill's account, however, and to do so he suggests we can understand the relation using Wilson's (2001) notion of "wide realization." The idea is supposed to capture properties that are contextually realized, that is, properties that only make sense in terms of a set of background conditions (Wilson suggests evolutionary fitness as an example (Wilson 2001, p. 13)). Wide realization means that a certain "non-core" part of the realizers is not in the entity (with evolutionary fitness,

this non-core part is the environment: a coyote is highly fit in the foothills of Los Angeles and significantly less so in the Amazon Basin). With "radically wide realization," the core part of the realizers is *not* in the entity.[12]

Schroeder suggests we can apply this to visual qualia: they are realized in the ventral stream in the wide sense (or radically wide sense, if you prefer), and the emergent is simply the non-core part (or core part) not realized in the brain. In this scheme, both the right patterns of activation in the ventral stream *and* the emergent must occur for qualia to be instantiated.

But doesn't this just undermine Megill's whole program? The idea was to get rid of realization altogether and this account relies on realization. Of course, the notion of realization is different here, but it comes to much the same. The emergentist says it's only *part* of the mind that's realized in the brain, but this doesn't do him any good. If we accept realization of any kind for our mental properties, it's hard to see how we can avoid collapsing back into property dualism, even if it is only partial property dualism, however we might make sense of that. A mental property is either realized or unrealized in the brain—you can't have it both ways. *But*, the emergentist objects, *realization as such isn't strictly what's important here, it's that this context obtains for this property to be instantiated*. So much the worse for the emergentist. Because then we're back to just a mere correlation between ventral stream activity and the visual quale, realized in the emergent. For wide realization to solve anything for the emergentist, he has to commit himself to the mental property being at least partially realized in the brain. But that's not an acceptable solution.

I think this is probably enough to show that Schroeder's recourse to wide realization doesn't work, but I'll note some problems that would arise even if the emergentist accepted what I said above and still tried to march on with this account. Let's say we accept that qualia are partially realized in the ventral stream, however that works out. Well which part of a quale is realized

there? Wilson's scheme applies to functional properties—we can make sense of how something like fitness is widely realized because fitness has a straightforward functional analysis: it's how likely the organism is to survive. In the simplest case this is a function of certain facts about the organism itself and its environment: i.e., there's a clear delineation between the core part (the organism) and the non-core part (environment). Qualia resist such an analysis because qualia are non-functional. What sense does it make to talk about the non-core part of a pain quale? Further, taking this line makes our emergents newly susceptible to causal impotence. If qualia are partially realized in the brain and partially not, then there must be some account for how the causal power is split up between the two parts. The emergentist can't deny that the realized part has causal power—we know very well that the synchronized firing of groups of neurons exerts causal power. But then what role is left for the non-realized part to play? If we want to locate the phenomenal aspects in either part, it seems like we must locate them in the emergent part. But then that sounds a whole lot like it's the non-causal/non-functional part that is realized in the emergent, in which case we're back to Lewtas's Dilemma. So wide realization, if it can be made sense of at all, makes the phenomenal aspects of qualia causally inert.

All this shows that wide realization is of no help to Megill. But Schroeder is correct in his diagnosis of the problem: without some sort of grounding, Megill's emergents seem to be physical in name only. Without any accountability to the physical world, it looks like we might be in danger of violating Closure after all. We can't just say the "mental" is physical unless it's physical in the sense of having such accountability. In this light, Megill's picture begins to look quite a bit like out-and-out ontological dualism, and this account is supposed to explain how emergent mental entities can be made sense of in a physicalist framework. In this, Megill fails and Schroeder's attempt to deal with the problem using wide realization does him no good.

Conclusion

Contemporary emergentist accounts of mind take pains to avoid causal exclusion while accommodating their theories to physicalism. This needle is just too difficult to thread, however. We saw that even if Zhong's account avoids causal exclusion by Kim's argument, the only way to make sense of his notion of "broadly physical" leads right into Lewtas's Dilemma and thus epiphenomenalism. Megill's emergent entities may initially do better on this count but without proper grounding—which wide realization cannot provide—his emergents don't really seem to be physical at all, in which case they can't be accommodated into physicalist ontology. Physicalist emergentism may have a certain initial appeal, but in the final analysis it simply cannot deliver on its promises.

Notes

1. Weak emergence is thus an epistemic doctrine. Notably, Chalmers cites the patterns that emerge in Conway's Game of Life as an example, something Dennett also uses when explicating his "mild realism" about patterns (Dennett, 1991). Terminology aside, the analyses are compatible with one another even though Dennett is an ardent eliminativist.
2. Depending on the type of emergence one postulates, one could be concerned alternatively with emergent phenomena broadly construed, or more specifically, with either emergent properties or emergent entities. Since the view in general is neutral with respect to these three uses, I'll use "emergent" to cover the general case and the more specific term when it's relevant.
3. Or whatever it emerges from: super advanced silicon-based computer, alien central nervous system, beer cans arranged such that they are a Turing Machine, etc. As we'll see, Zhong isn't committed to neural chauvinism, though Megill might be, for reasons to be discussed below.
4. I'm glossing over the details here. Kim's model for reduction involves 3 steps: first, functionalizing the higher-order property; then, identifying realizers of the higher-order property in the base; and finally finding a base-level theory to explain how the realizers "perform the causal task" of the higher-order property. See Kim (1999), pp. 9-13.
5. One may object that M1 should actually be considered an intention to drink from either the left or right bottle. I don't think this is right, but I also think it wouldn't matter if it were. If you prefer, imagine two cases in which I have

the same intention to drink from the left bottle: in one case a few neurons aren't firing while in the other they are. The difference of a few neurons is not going to be enough to change my intention, though the two cases would be different physical states, i.e., P1 would be different.

6. Or, more precisely, it's an essential *part* of the cause of P2. Zhong doesn't think M1 alone is sufficient for causing P2 and in the counterfactual where M1 obtained but P1 didn't, P2 would not be caused. M1 and P1 are thus together the cause of P2. While Kim ruled out partial causation because of concerns about overdetermination, or violation of Closure, with Closure* this isn't a concern, because M1 is physically-grounded.
7. Megill's account (considered in section 5 below) also doesn't rely on passive causation. And either way, passive causation doesn't work for any emergentist theory of mind. See Lewtas (2017), pp. 480-484 for a full discussion.
8. "[T]hese magnetic fields are clearly not reducible to their base; one cannot reduce a magnetic field created by neurons to the neurons themselves." Megill (2015), p. 611
9. The evidence primarily comes from Transcranial Magnetic Stimulation (TMS) studies.
10. Technically, Kim asserts that it isn't that P causes M so much as M is an emergent property that supervenes on P. Since strong supervenience is out, it makes more sense to construe the relation between emergent and base as causal. Megill is still committed to *global supervenience*, however, which he (rightly, I think) considers necessary for minimal physicalism.
11. That being said, I do think Megill would also have to contend with the standard objections leveled against identity theories: EM fields are not red or painful and it's hard to imagine a physical object—emergent or not—that would have such qualities. I'm bracketing this (valid) concern because I think the grounding problem is fatal anyway.
12. Wilson suggests social actions as example. See Wilson (2001), pp. 13-14

Bibliography

Chalmers, David J. (2006) "Strong and Weak Emergence," in In P. Davies & P. Clayton (Eds.), *The Re-emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, pp. 244-246 (Oxford: Oxford University Press)

Dennett, Daniel C. (1991) "Real Patterns," *The Journal of Philosophy* 88(1), pp. 27-51

Persons Everywhere Everywhen and One of Each: Motivating Toward Observer Presentism Panpsychism by Answering the Transitive Objection (or Everyone Gets a Present)

Daniel Castro

Introduction

Lots of philosophers are uncomfortable with consciousness, think it is something to be done away with. It's kind of like the ontological question, *"Why is there anything at all?"* What an annoying question. Many philosophers prefer to just put an end to the question—saying things like, *"That's a meaningless question; it makes no sense to ask that."* In this way many philosophers try to get rid of consciousness: it is not *really* what we think it is; it does not really exist. It is something to be eliminated, thrown away. Why don't we dispense with water and air too, while we are at it?

Alas, we never stop asking, *"Why?"*

Let's get to some definitions then.

In the philosophy of time, there are opposing views called "presentism" and "eternalism." Eternalism is the view that all times (and all the different things in existence at those different times, e.g., the younger version of you and last night's dinner) are equally as real as everything that is real in the present moment, the "here" and "now" (let that sink in). Presentism, on the other hand, can be looked at in at least two ways: what it says about time and what it says about ontology (the being of objects). I will be focusing on what it says about ontology. As a favor to you, so as to not keep you in suspense, and so we can move on without a cliffhanger situation, I'll tell you what presentism says about

the time question. Here it goes: presentism entails that *only the present moment exists*—no other times exist like they do in eternalism, in which every time exists, all the time. The end. Onwards to the ontology question.

Presentism says that no objects exist, except what exists in the present moment; some philosophers see moments as "abstract representations of instantaneous states of the world" (Crisp 2007, p. 40). This presents a problem: the special theory of relativity (STR) goes against presentism. STR entails the relativity of simultaneity (RS); RS means there is no absolute "now"; so, depending on frame of reference, an event E1 can be "happening now" in one frame of reference, and another event E2 can be happening before or after E1, but it is still "happening now" in its own frame of reference. We can see the possibility creeping in that one thing exists in a substantially real way at one relativized "now" and yet exists-not in a substantially real way in a *different* relativized "now." In STR, there is no privileged present, nor special ontological status for any objects in the universe. It seems STR and presentism are incompatible, and STR is a widely accepted scientific theory. Too bad for presentism. Unless…

Observer presentism is the view that for every observing-capable entity in a reference plane, there is a set of objects which exist and nothing else exists for that observer and no other times exist at that reference plane. I defend observer presentism against the transitive argument (explored below). If my argument against the transitive objection succeeds, then observer presentism may be compatible with STR.

Panpsychism is essential to motivating observer presentism and answering the transitive objection. Panpsychism is the view that consciousness is fundamental (like fundamental particles and laws) and that consciousness is widespread, not just limited to biological systems (there are different panpsychisms, but this is what I'll mean by panpsychism). I won't be defending panpsychism here. I'll simply be assuming panpsychism to show observer

presentism can overcome the transitive objection.

My plan is simple. The transitive objection causes trouble for presentism, if STR is true. Presentist panpsychism motivates a move towards observer presentism (OP), which then surmounts the contradiction raised by the transitive objection and lives in harmony with STR.

Section I: I explain the problem between presentism and STR. Section II: I define observer presentism and describe panpsychism. Section III: I go over the transitive objection against observer presentism; one must choose either presentism or STR (and right now that doesn't amount to much of a choice). Spoiler alert: STR wins because of science. Section IV: I show how combining panpsychism with presentism can save observer presentism from the transitive objection; thus allowing presentism and STR to live in wedded bliss. Let's get on with it.

Section I: What Are You Talking About? What's the Problem?

STR tells us that there is no privileged present. An analogy to this would be that there is absolute space. As Newton said, "Absolute space, in its own nature, without regard to anything external, remains always similar and immovable." This is the idea that there is a space that makes up the structure of space, never moving—no relative space. Is there also absolute time and therefore no relative time? Is there an absolute present? Einstein says no. STR says that there are times that exist for some observers and yet are past or future to other observers. Therefore, some things exist and don't exist, depending on the frame of reference. This is fully compatible with eternalism.

Presentism says only the present moment exists and only present things exist. It says that time is absolute, no matter the frame of reference. One of Einstein's main ideas concerning STR is that "there is no privileged foliation of spacetime (or our models

thereof), no foliation such that it alone tells the correct story of how the cosmos evolves over time" (Crisp 2007, p. 4).

Can presentism and STR both be true in a robust way? It doesn't seem like it, but I will contend that they can be, while still maintaining the spirit of both theories.

Section II: What Are Observer Presentism and Panpsychism?

There are different ways of articulating presentism and panpsychism and different aspects to emphasize. In essence, classical presentism says that only the present moment is real, and only the things that exist at that moment are real. Furthermore, the present moment, and the things that exist at that moment, are privileged, and no other times or things possess a concrete reality. Observer presentism (*O-presentism*) says that what exists is relative to each person/observer, and for each observer, only the present moment exists—the only objects that exist are those that exist at that moment, nothing before nor after is concretely real.

Panpsychism says that consciousness is fundamental and widespread in nature. Some panpsychists maintain that all fundamental entities are conscious, others maintain that only *some* fundamental entities are conscious. They both agree that consciousness is fundamental, the way other entities are fundamental (e.g., natural laws, subatomic particles). If a thing is fundamental, it is not explainable in virtue of, or in terms of, something else. It just is.

Section III: The Transitive Objection

The transitive objection was introduced by Hilary Putnam (1967). In essence: if Object A is real to Object B, and Object C is real to Object B, then Object A is real to Object C. The "real to" relation here would be referring to objects that are (at any given moment(s)) occupants of the same inertial frame of reference. Putnam used

this to describe what reality should look like for eternalists. The conclusion is that contemporary theories entail that both past and future things are as real as present things, because of different planes of reference—and of course this works fine for eternalism.

Mark Balaguer (2021), Steven F. Savitt (2000), and Mark Hinchliff (1996) have used the transitive objection to argue against observer presentism. It is not an argument against presentism per se nor an example of how presentism and STR are inharmonious, but an argument against *O*-presentism. The problem they see is that if O-presentism is true, then it is contradictory. Here I tweak Balaguer's presentation of the transitivity objection to O-presentism:

> *The Transitivity Objection to O-presentism:* At the moment of an encounter between two people (e.g., when P1 flies past P2 in a spaceship), both people are real to each other. Thus, since the "real-for" relation is transitive (it's a relation of coexistence), it follows that whatever is real for P1 is real for P2. But O-presentism entails that certain things are *not* real for P1 that *are* real for P2 at the moment of our encounter, so the view is false.

As you can see, there is a problem. Above, I said O-presentism says that what exists is relative to each person and/or observer, but if the transitive objection is right, we get:

(1) According to O-presentism, whatever exists is relative to each person, and only the present moment and present things exist.

(2) If P1 is flying super-fast (like super-fast and therefore in a different plane of reference) past P2 then there are certain things that are real for P1 that are not real for P2, because there are relative momentary realities per person, so says STR.

(3) The "real-for" relation is transitive—relation of coexis-

tence—so if P2 is real to P1, and P1 is real to P2, then whatever is real to P1 is real to P2. Logically:Object 1 (O1) is real to P1. P1 is real to P2. Thus, O1 is real to P2.

∴ The moment P1 and P2 meet, on P1's extremely fast journey, there are certain things that are both real and not real to P2, at the same time.

No bueno: a contradiction for O-presentism. This is the transitive objection to O-presentism.

Section IV: Observer Presentism Panpsychism

There is good news for O-presentism, if we assume panpsychism. Panpsychism says we have potentially infinite fundamental and conscious physical entities widespread in nature. O-presentism says that what exists is relative to each observer (each of those infinite conscious entities); for each of those realities, the only things that exist are objects that exist presently.

If we put panpsychism and O-presentism together (OPP), we get an infinite amount of personal, privileged, present realities (sets of objects per conscious entity, perhaps abstract); and it is real presentism, even though it means there are potentially infinite presents. It is no less presentism for each conscious entity just because they don't share one grand presentism for all.

What we have here is a set of possible worlds within our actual world. Not quite Lewisian possible worlds: A difference between Lewis's possible worlds and all of these actual worlds (the worlds of each conscious entity) is that these worlds *do* collide and intersect.

Finally, if every conscious entity has its own reality (its own possible world), then transitivity doesn't even apply. Think of P1 flying by P2 super-fast. We actually have two worlds here, not one shared one. For P1, a wholly private conscious experiencer with her own relative reality, P2 and the objects in its plane of reference

are real. For P2, P1 and the objects in her own plane or reference are real. Adding panpsychism, we get a presentism that is also relativistic. The panpsychism connects the O-presentism to STR and the relativity of reality. If we think of OPP in the right way, true to what it is *actually* proposing, there is no reason to think that transitivity would come into play—just like transitivity would *not* come into play between Lewis's possible worlds.

I, Conclude

A lot of philosophers have been trying to get rid of consciousness, but we shouldn't try to get rid of it. Instead, let's put it *everywhere*, fundamental in nature. This brings about the harmony between our intuition that the present is privileged and that Einstein's relativity is true. There are many presentists and panpsychists making good arguments for their views. I leave that to them. Here I have simply answered the transitive objection, showing that OPP is a good way forward for presentism. So, if you down with OPP, yeah you know me.

Bibliography

Balaguer, Mark. (2021) *Metaphysics, Sophistry, and Illusion* (Oxford: Oxford University Press)

Crisp, Thomas. (2007) "Presentism, Eternalism and Relativity Physics," in: W. L. Craig & Smith (Eds), *Einstein, Relativity, and Absolute Simultaneity*, pp. 262-278 (London: Routledge)

Hinchliff, Mark. (1996) "The Puzzle of Change," in: J. Tomberlin (Ed), *Philosophical Perspectives* 10, *Metaphysics*, pp. 119-136 (Cambridge, MA: Blackwell Publishers)

Putnam, Hilary. (1967) "Time and Physical Geometry," *Journal of Philosophy* 64: 240-47

Savitt, Steven. (2000) "There's No Time Like the Present (In Minkowski Spacetime)," *Philosophy of Science*, 67, Supplement, pp. S563-S574

Zimmerman, Dean. (2011) "Presentism and the Space-Time Manifold," in: Craig Callender (Ed), *The Oxford Handbook of Time*, pp. 163-244 (Oxford: Oxford University Press)

The Right to Choose a Deaf Child

Alexandra Meyering

Introduction

With the advent of preimplantation genetic diagnosis (PGD), an array of options for prospective parents has become available. While it is more typical for parents to favor embryos that will be born as hearing children, many deaf couples also desire to have children who share their condition.[1] While many in mainstream society would balk at such a decision, we must ask ourselves the following question: how much of the common uneasiness felt at intentionally selecting for deafness occurs because of a negative "master narrative" that dictates that disabled lives are not as full or rich as abled ones? This attitude arises from a deeply embedded societal meta-narrative that has taken shape over centuries—one that ultimately values the expectations and intuitions of abled people more highly than their counterparts. I hope to challenge some of the assumptions that some may take for granted due to the pervasive nature of the master narrative, focusing specifically on the topic of deafness and reproductive bioethics.

In our society, the notion that parents should be legally able to raise their children in a way they see as serving their best interests is uncontroversial—that is, as long as no direct harm is brought to the child or others as a result, and that the child's future opportunities for a good life are adequately preserved. Societies that feature liberal neutrality as part of their foundational principles uphold a view that it is fair to honor diverse interpretations of what makes a good life. Another one of these tenets is the commitment to protecting everyone's right to pursue happiness in a way that suits them best. To stay consistent with these principles, one's views must avoid creating ableist[2] double standards.

In this paper, I will argue that intentionally selecting a deaf embryo should not be considered a harm, and that there is no good reason within the boundaries of liberal neutrality to single deaf parents out from others who use PGD. I will do this by showing how selecting deafness for one's child does not violate their right to a sufficiently open future, as detailed by Joel Feinberg (1980). This will be accomplished by first addressing three arguments against selecting for deafness. I will then discuss how the master narrative distorts common intuitions about disabled lives, leading to the belief that selecting for deafness is a harm. To conclude, I will offer three arguments in favor of allowing for deafness selection. This paper will aim to show that deaf children have ample opportunities for thriving, that their impairment does not necessarily preclude a good life, and that it is morally permissible to intentionally select a deaf embryo.

Section I: Arguments Opposing Deafness Selection

To begin, I will examine some initial objections to choosing deafness for one's child. Three arguments commonly used to oppose the intentional selection of deafness are as follows: (a) it would significantly limit the child's lifetime opportunities, (b) it is morally impermissible to deafen a hearing infant, and as such we should not select a deaf embryo, and (c) since being deaf comes with significant societal disadvantages, it is wrong to knowingly force a child into such a position.[3] I will address each of these positions in turn and demonstrate why these arguments do not provide good enough reasons to deem selecting deafness as a morally impermissible act.

1.1: The Preservation of an Open Future

Arguments that oppose deafness selection via PGD oftentimes discuss the concept of preserving the "child's right to an open

future," as it is defined by Joel Feinberg. In his 1980 essay, Feinberg argues that we ought to protect a child's right to enjoy a future that is full of as many opportunities and options as reasonably possible. Feinberg also states that parents should protect a child's "rights in trust"—the rights a child will be entitled to in adulthood, but cannot exercise before maturity (Feinberg 1980). An oft-cited example of a 'right in trust' is the right for a small child not to be sterilized so that they may one day have the option of choosing to become parents themselves, even though it is not a possibility in youth (Davis 2012).

As applied to the topic of deafness selection, those opposed believe that if one prevents the child's ability to hear (and thus the opportunities that come with that sense), then the child's future educational and career choices, recreational abilities, marriage options, and social sphere are significantly narrowed (Davis 2012). In this view, being deaf automatically closes off such a sizable pool of opportunities to the child that their futures will inevitably become unreasonably limited. They argue that it is, therefore, a violation against the child's rights in trust. By this line of reasoning, the violation of these rights leads to the child's inability to enjoy the same quality of life as a hearing child, thus depriving them of a significant amount of autonomy.

However, simply having a wide range of opportunities or a large degree of autonomy available does not necessarily ensure a good life, as argued by Cristian Puga-Gonzalez. He points out that it is not consistent with the principles of liberal neutrality to consider autonomy as *the* determining factor of a valuable life. He reasons, "if one appeals to autonomy to prevent [a deaf child's] existence, one would be claiming that the state can rightfully prevent the existence of persons with lower degrees of autonomy, as if a higher degree of autonomy, by itself, could make lives more valuable (as if autonomy was a constituent element of 'the good')" (Puga-Gonzalez, p. 374). To be truly fair to all reasonable conceptions of 'the good', we cannot use a slightly less-than-average

scope of autonomy as a marker for whether a life is worth living.

Additionally, it doesn't necessarily follow that having a large number of future options available will lead to a higher quality of life. A smaller number of available options does not prohibit enjoying a fully meaningful existence. J.M. Wallis (2020) suggests that there is evidence to show that fewer options can sometimes produce more overall satisfaction:

> More abstractly, results in cognitive science do not indicate that increased opportunities correlate linearly with increased well-being. Excessive multiplicity of choice can cause anxiety or be overwhelming. Robert Sparrow (2002, p. 11) raises a related point in "Better off Deaf" when he observes, "[i]n a society which fetishizes individual choice and opportunity, it may seem obvious that these are goods." Perhaps more options only increase our happiness to a point, and after that point, increasing options are neutral with regards to happiness. Without solid evidence, however, we ought not insist that the reduced magnitude of future openness for a deaf child could be grounds to make selection for deafness necessarily impermissible. (Wallis, p. 7)

To build on this, take into consideration the fact that any parenting style or life situation whatsoever necessarily limits the number of options available to a child, and so it is unfair to distinguish deafness as a singularly limiting quality. Stramondo (2020) argues that the conditioning of any lifestyle or selection of any body, abled or otherwise, will inevitably close off a large number of options. The way a child is raised in terms of values, religion, personal freedom, location, or any number of other factors is going to form much of their worldview and eliminate certain possibilities. Additionally, a child's available opportunities will be altered drastically no matter what sort of body is chosen. On this subject, Stramondo writes:

> If a parent chooses any such body for their child, then they are inevitably also choosing a habitus for that child that will narrow their range of autonomous choices by producing a variation on the "system of tacit rules governing practices and behaviors." [Scully, 2008] Even more significantly, for our purposes, this habitus is not subject to reflective revision later in life, and so, simply put, one does not choose their own habitus. (Stramondo, p. S32)

Many would intuitively accept the decisions of abled parents to choose a habitus on behalf of their children that will significantly alter the number of future opportunities available to them, especially if it is of great cultural or religious significance to them.

For example, consider the situation of a language that is spoken natively by a comparatively few number of people. To the parents, it may be important to raise their child to speak the minority language as their native tongue in order to preserve an important part of their culture. Most would not see this decision as impinging on the child's rights in trust or right to an open future. If the parents teach the child the dominant language first, this would certainly give them more opportunities. The child may also not be able to communicate as fluently in the dominant language because of their delayed exposure to it. This is likely a choice that will limit the child's future opportunities to some extent. However, it is still a fair and understandable decision to raise a child in a way that protects a minority culture, and most would not object to this. It is natural to recognize the value in preserving a minority culture whose existence may be already threatened, and so the same understanding should be afforded to the deaf culture.

Another example of an intuitively morally permissible parenting plan that will inevitably reduce a child's future opportunities to some degree is the choice to raise a child in a very strict religious environment, such as that of the Amish, Jehovah's Witnesses, or Hasidic Judaism.[4] Liberally neutral societies recog-

nize the value of religious liberty, even in a profound form, as part of their commitment to preserving diversity and the freedom to choose a preferred lifestyle (Puga-Gonzalez, 2019).[5] It is clear that to raise a child in a strict religious environment significantly limits their choices, oftentimes permanently because of the deep impression these environments often make on children. However, it is still commonly held that this choice should be honored as part of a parent's right to choose what will bring the most well-being to their children's lives. If it is fair to say that parents should have this right when it comes to religious beliefs and linguistic decisions, then, to remain consistent, there should not be a double standard held against the deaf community. If it is agreed that an Amish child, a child who speaks an obscure language, or even a child who was raised in a remote location far from large cities still has an adequately open future, then an intentionally selected deaf child does, as well. With all these points taken into consideration, it is clear that deafness alone does not dramatically infringe on a child's right to an open future, and should not be disallowed on those grounds.

1.2: If it is Wrong to Deafen an Infant, then it is Wrong to Select a Deaf Embryo

All would agree that it is obviously morally impermissible to deafen a hearing infant. Many opposed to deafness selection would use this fact to argue that, likewise, it should be intuitively morally impermissible to select a deaf embryo. Most would consider it a wrong to intentionally remove a baby's sense faculty. Some, then, posit that selecting a deaf embryo is akin to depriving a child of a sense in much of the same way, and therefore it is wrong to do so.

However, this argument is flawed in that it relies on a weak analogy; deafening an infant and selecting a deaf embryo are not morally equivalent or even analogous actions. Deafening an infant deprives a specific individual of a sense. This is a deliberate and

dramatic act of harm done to a living person. These sorts of actions are regarded as moral harms against living individuals (Wallis, 2020). In the situation of a deafened infant, a living person, who *could* have been hearing, is robbed of an existing faculty; and this is a transgression.

In contrast, selecting a deaf embryo brings a life into the world that could *only* have been deaf. In this situation, it is not correct to assume that the *same* child could have been born with the ability to hear, and that the parents are depriving a certain individual of that sense. Before implantation, the child in question is still a non-person—what we are thinking of is not a person, but rather the *potential* for there to be any child at all.[6] Selecting a deaf embryo cannot do harm to a child because no such child exists yet. (Häyry, 2004) By choosing a deaf embryo, the parents are bringing to life a single individual whose only possible fate was deafness. Because no harm is done to an existing person by taking this course of action, the argument that deafening an infant is morally equivalent to selecting a deaf embryo does not hold.

1.3: The Argument from Avoiding Disadvantage

Some commentors take the position that people ought to refrain from having children who will experience societal disadvantage because it is unfair to the child. These people may say that even if we grant that some opportunities in the deaf culture are closed to hearing people, there are significantly more opportunities that are closed to deaf children in the hearing world. As such, it is blameworthy to bring a child into the world who cannot access those opportunities, specifically because it puts those children at a disadvantage in life. Furthermore, while those that take this view recognize that there is innate value to the lives of deaf people, they also argue that society, as it is now, will still treat deaf folks unfairly, so it is wrong to have a child who will inevitably experience those unfair conditions. They argue that they are putting the

child into an unfairly difficult position in life that could have been avoided had they not been born.

However, this same argument doesn't seem as convincing if it is contrasted to other everyday reproductive decisions that most will readily grant are morally permissible. For example, though women may be at a disadvantage in our current society, it is not considered wrong to select a female embryo over a male one because of the challenges the child will face.

Outside of the realm of selecting for traits via PGD, it is also not thought of as wrong to have a child who will inevitably experience disadvantage in some form. Consider the disparity of fair treatment between white people and non-white people in the USA. People of color are put into a disadvantaged position because of a racist and prejudiced dominant culture. Much like disabled individuals, non-white people will face significant difficulties due to the moral failings of a majority culture that does not treat them with equity. However, no reasonable person should feel comfortable making the argument that non-whites should refrain from having children altogether because their children will face disadvantages that others will not. That would be a preposterous claim that we would easily recognize as racist (Davis, 2012).[7]

A final example that reveals the unfairness of the "argument from avoiding disadvantage" is the decisions of economically struggling families to have children. While it must be acknowledged that those with fewer financial resources are at a disadvantage in society, most would find it abhorrent to say that a non-wealthy family should not be allowed to reproduce because they are unable to provide every possible advantage to their potential children. Society upholds that their decision to have children should be legally permissible, no matter their level of economic prosperity.[8] Even though that child will most likely not have access to as many opportunities as a child born to a wealthy family, it is still ultimately a right that a parent ought to be able to exercise.

With all of this in mind, it's fair to say that it is intuitively

permissible to bring children into the world who are disadvantaged by society; most do not hold that only those parents who can provide every possible advantage for their children should be allowed to reproduce. As such, there is no suitable reason to single out deaf people in this respect, if one is to remain consistent with this view.

Section II: Dismantling the Master Narrative

Before I can offer arguments in favor of deafness selection, I must address two larger, related views that are associated with the topic: the mistaken idea that disabled lives are somehow of inherently lesser value than abled ones (the master narrative), and the idea that because of this, selecting for deafness is a harm.

Master narratives appear in many forms. The long-held but obviously sexist belief that women should stay home, take care of children, and serve their husbands is an example of how these types of meta-narratives manifest in our world. This specific version can be traced back even further to when women were thought of as property on nearly a global scale. This variant of the master narrative casts men as superior to women, even though there is no reasonable support for that claim. However, its influence and hold over society is undeniable.

Much like this, a master narrative constructed against people with disabilities exists in our society. We have seen this narrative function to produce plenty of morally outrageous actions throughout history, such as the practice of leaving a disabled infant out to die or the genocidal eugenics movement of the 20th century. Though these particular offences are now widely condemned, it is still possible to see the influence of the history of ableism reflected in our daily lives. Think of wheelchair inaccessible sidewalks or bathrooms; elevators without braille; slurs that denigrate people with cognitive disabilities; the idea that prescription drugs could

somehow alter the "true self" of a clinically depressed person, or that they are reserved only for the weak-willed; the practice of aborting fetuses with Down syndrome; the reality that people with mental illness are routinely taken away in handcuffs and treated as criminals during mental health crises; television programs that present disabled people as either sources of pity or inspiration meant for the abled population; and the way mental illness is demonized or mocked in our pop culture. These are but a few common examples in an undeniably oppressive society.

Generally, the ableist master narrative usually boils down to the idea that disability is an inherently negative thing, and that living a disabled life is, in some way, never as good as living an abled one. When people attempt to prevent disabled children from being born or when disabled lives are put to an end out of so-called "mercy," it has been cast by this narrative as a choice made for the "greater good." The narrative transmits the idea that it is better to deprive disabled people of their lives entirely rather than to let them live with an unusual condition or with the loss of a faculty. This is the evidence of the insidious and domineering influence of the master narrative.

Stramondo seeks to explain the origin of these mistaken intuitions and the master narrative they form. He cites the work of Scully to show that these kinds of meta-narratives arise from the collective habitus of the majority. A habitus might be defined as the seemingly naturally arising views, behaviors, and intuitions that become crystallized in a person's worldview because of the way they've lived and the groups they've largely interacted with. (Stramondo 2020) For example, a seeing person may not even notice that a busy streetlight lacks an audio signal to let blind people know when it is safe to cross because the experience of blindness is not part of his habitus. Similarly, the same seeing man may regard the blind person's habitus as something inherently undesirable because of how it differs from his own.

Stramondo shows that the way disability is perceived by

society at large is part of this ingrained, collective, majority-built habitus, and that many of the widespread assumptions it shares are due to the stories and intuitions that come embedded within it:

> Identities are constituted from the first-person perspective through the loosely connected stories we weave around the things about us that matter most to us: the acts, experiences, and characteristics we care most about, and the roles, relationships, and values to which we are most deeply committed. In the course of this narrative construction, we draw on stock plots and character types that we borrow from the familiar stories embodying our culture's socially shared understandings—the stories that I have been calling master narratives. (Stramondo, p. S35)

Within the abled-centric view's habitus, these stories have created a master narrative that conveys that it is intuitively better not to be disabled, and that any sensible person could not intentionally choose a disabled life for their children. However, Stramondo points out that the habitus of a disabled person vastly differs from the majority's, reporting that deaf parents will often say that it is a rational, natural, and justifiable attitude to want to have a child who shares their condition. To assume deaf parents' intuitions and judgments are, on the whole, flawed and incorrect, is the influence of the ableist master narrative: "…we can't just assume that any and every disabled person's habitus has given them suboptimal preferences that interfere with their autonomy. There would need to be an independent argument for why the desires and values embedded in a nondisabled habitus are more desirable and valuable." (Stramondo, p. S33)

To assume that deaf people do not know what is best for their children, or are somehow mistaken about the true value of their lives is an ableist notion that arises from a society that negatively and unfairly views disability in general. Furthermore, it is a dangerously eugenic attitude to think that society should disallow

the births of children whose conditions may seem undesirable in the mainstream's conception of what makes a good life.[9]

Additionally, Stramondo shows that the master narrative inflicts double damage in that (a) it tells a false story about the value of disabled lives, which leads to fewer resources, and less awareness and dignity for disabled people, and (b) is hypocritical in that it effectively works to limit the open futures of disabled children by its influence. Stramondo argues that the internalized ableism that comes from the master narrative erodes the sense of worth of disabled people, and serves to reinforce already imbalanced power structures. He shows that by constantly telling a story to a person that communicates that their life is of lesser value or contains fewer opportunities, they will come to internalize this message, at least partially, and their future opportunities will diminish as a result. He calls this phenomenon infiltrated consciousness. He believes that if we truly cared about providing everyone with an adequately open future, then dismantling this narrative and allowing for more inclusion would be more effective than disallowing the selection of deaf embryos.

With this discussion in mind, it is reasonable to assume that a large part of the common, kneejerk intuition that deafness should not be intentionally selected is fueled by the ableist master narrative that has permeated the dominant worldview. In the hopes of fairly evaluating the issue, allow the following discussion to move forward with the willingness to dismantle the master narrative in mind.

2.1: Selecting for Deafness is Not a Harm

Another salient aspect of the master narrative is the idea that it somehow would harm a child to bring it into the world while knowing that the child's impairment will produce perceived suffering of some sort. But can bringing a disabled child into the world accurately be said to produce harm in any coherent

sense, or is this also a distortion created by the dominant cultural atmosphere?

To show how it is not actually possible to harm a child by selecting a deaf embryo, I will briefly return to a subject that I touched on in Section 1.2. Crucially, it must be recognized that the embryos in question have yet to become people. In the process of PGD, a number of viable embryos are presented to a parent, who can then select which one they wish to carry to term. The loss of the unselected or discarded embryos is not considered the loss of a "person." Therefore, it's possible to consistently argue that no people exist prior to implantation, and that no harm has been done to any one person at this point. This line of thought is generally known as the nonidentity problem, which was made famous by the work of Derek Parfit. Puga-Gonzalez details how selecting a deaf embryo cannot feasibly cause harm:

> On a comparative account of harm, an action performed at time t1 is harmful to a person P, if and only if such action causes P to be worse off at some time latter t2 than she was at t1 (Williams and Harris [48], 344). By arguing that [the child] was harmed, therefore, one would be claiming that implanting the embryo that became [the child] caused [her] to be worse off (existing being deaf) than she was before being implanted (non-existence). But this claim would present a logical impossibility: one cannot make any well-being comparisons with non-existence because 'non-existence is no state at all' (Williams and Harris [48], 345). 'Since it is necessary to be if one is to be better off, it is a logical contradiction to say that someone could be better off though not in existence' (Feinberg [16], 16). Hence, it is not logically possible to claim that [the child] was harmed on a comparative account. (Puga-Gonzalez, p. 368)

With this in mind, it is reasonable to dismiss the idea that selecting for deafness violates the well-being of a child to a degree

that can definitively be considered harm. If the idea that something must first exist in order to suffer harm is acceptable as true, then it must be conceded that no child is harmed by the selection of any given embryo.

Section III: Arguments in Favor of Deatness Selection

I have attempted to illustrate that some of the commonly used arguments that oppose the selection of deaf embryos are unfair if one wishes to remain consistent with many of the values often held surrounding reproductive autonomy—one cannot impartially use these complaints as good reasons to declare deafness selection morally impermissible. Moving forward, I will focus on some reasons that support the selection of deaf embryos. These views show that intentional deafness selection allows for flourishing in a child's life, and that it is comparable to the lifestyles often enjoyed by hearing individuals. Though there are many such arguments, I will focus on three that I find most compelling. They are as follows: (1) Deafness is a necessary condition for fully living in deaf culture, (2) one should stay committed to honoring reproductive autonomy for all members of society, and (3) deafness can reasonably and justifiably be considered an enhancement in certain aspects of life.

3.1: Cultural Immersion as a Vital Element for Flourishing

While a hearing child brought up in a deaf family would certainly gain great insight into both the hearing and deaf worlds, it is arguable that some of the subtleties and nuances of the culture cannot fully be accessed by a hearing child. It is the concern of deaf families that their hearing child might not be able to fully experience the extent of the richness of deaf culture, art, and language. They also worry the child may experience a loss of comradery and sense

of belonging due to not sharing the same condition as one's family and close friends.

A helpful analogy may be found by imagining a dynamic between a pair of siblings, one of whom lives with a pronounced form of projective synesthesia, a condition that will likely dramatically change the one sibling's perception of the world.[10] The neurotypical sibling may gain a deep sense of understanding of the neurodivergent sibling's experiences due to their proximity and emotional closeness. They might be able to gather more insight into how their sibling sees the world. However, they cannot fully know what it is like to live with synesthesia, and to be able to taste sounds, for example. They are not completely privy to their sibling's worldview, but rather only to the descriptions of it.

According to some deaf families, the loss of the condition of deafness would deprive their future child of some very important, enriching life experiences and qualia, and thus the child could not fully flourish according to the parents' value system. These families so highly value the benefits of being deaf in deaf culture that they view the condition as necessary for maximizing all the child's best interests. While this may be at odds with a majority view, if one wants to remain consistent with the commitment to valuing diversity in the opinions and lifestyles of all people in our society, then one must allow for this interpretation of flourishing.

3.2: Honoring Reproductive Autonomy

In most liberal societies, all people are (in theory) entitled to reproductive rights and autonomy without exception. It is also assumed that the state should remain impartial to its citizens' value systems, unless they are unacceptably intolerant or harmful. This is the essential idea behind the theory of liberal neutrality. Puga-Gonzalez writes, "The general idea of liberal neutrality is that the state should not justify its fundamental political principles and institutions, or laws, by appealing to the superiority or infe-

riority of any particular conception of the good or by showing a preference for it on some other basis." (Puga-Gonzalez, p. 364)

As such, one of the tenets of a liberally neutral society is that almost all visions of "the good life" must be allowed to be pursued, unless it causes serious harm or injury to others or oneself. Under these rules, most agree that all people are entitled to make decisions regarding their futures as parents, including how to raise their children with their best interests in mind; there are few situations imaginable where it would be justified to withhold these rights from prospective parents. So, to remain consistent with the principle that all people in a society should be extended the same options, protections, and rights regarding reproductive autonomy, it would be wrong to single out those with disabilities from all others on the grounds that their lives don't seem worth living to some people. To restrict the options of deaf couples or individuals based on mainstream society's arbitrary value system is a violation of these basic reproductive rights.

A response to this argument may be that if one is concerned with protecting the autonomy and rights of all people, then not only the parents' right should be honored, but also those of the future child. This view maintains that people should protect all the rights in trust for children who have yet no say in the matter. Those who hold this view believe that society must accept the prohibition and violation of some of the parents' reproductive rights in order to preserve the rights of the unborn child.

However, it seems reasonable to suggest that the rights of currently living persons should be treated with some precedence over the rights of unconceived, potential persons. While the rights in trust and the open futures of potential persons should be honored, at the time of the decision to select an embryo, the child still does not yet exist. Most would intuitively give more respect to living people rather than to potential people. In the case that the parents would be planning to do egregious, unnecessary harm to a potential person, then this assumption should be quali-

fied to account for that. But, as has been previously established, deaf children have access to an adequately open future that has a reasonable number of opportunities for flourishing. Deafness alone does not cause egregious or unnecessary harm. In the case that the potential child will not be subjected to profound harm, one ought to place the rights of the currently existing parents in a position of higher importance than that of a child who does not yet exist.

3.3: Deafness as an Enhancement or Benefit

It is not unusual for people in the deaf community to see their condition as something preferable to hearing—something that enhances the quality of their life. There are plenty of deaf people who do not wish to identify as disabled, but rather as a member of a distinct culture. In her 2019 essay, J. M. Wallis identifies many aspects of deaf life that are often considered positive enhancements.

The first of these positive aspects is the unique nature of sign language, which allows for a strong visual component of communication—Wallis quotes Sacks as calling it "spatial and cinematic" (Wallis, p. 6). An intimate understanding of this language seen from the deaf perspective could have obvious benefits in the artistic as well as linguistic realms. There are other clear advantages to sign language, such as the ability to communicate over a distance and its similarity to other regional dialects, offering a more accessible way to communicate across cultures. Another benefit Wallis discusses is deafness's consequent honing of visual and vibrational senses—she brings up the case of a deaf woman who can identify musical patterns such as fifths from the type of vibration the sounds make.

Wallis also mentions the idea that sometimes having less sense stimulation can be a benefit, because it can allow for deeper focus on other senses. She highlights the case of Kitty O'Neil, a highly successful deaf stunt woman and high-speed driver who

not only broke a land speed record, but exceeded it by over 200 miles-per-hour. In her statements, O'Neil attributes her deafness as a helpful factor that kept her from being overwhelmed while driving and allowed her to achieve her goals with greater ease.

In addition to these benefits, a large number of deaf people report that they are perfectly happy as they are and have no desire to change their condition. This alone is a type of benefit, in and of itself, and Davis quotes Roslyn Rosen, as saying, "I'm happy with who I am, and I don't want to be 'fixed.' Would an Italian-American rather be a WASP? In our society everyone agrees that whites have an easier time than Blacks. But do you think a Black person would undergo operations to become white?" (Davis 1997). One can reasonably argue that contentment is a good thing, and thus is an aspect of deaf life that obviously may be considered a benefit.

If, once again, one wishes to stay consistent with the values of liberal neutrality and allow for multiple interpretations of the good life, one must concede that there are enough good reasons given by deaf people to consider deafness as a type of enhancement. Most would readily accept any other parent wishing to enhance their child's life with their best interests in mind. There, then, is also support for selecting for deafness on these grounds.

Section IV: Moral Implications

The purpose of my argument in this paper was to highlight a double standard in bioethical debates that negatively affects disabled individuals and perpetrates a deleterious master narrative. I intended to show that society should not allow its unconscious biases to prevent disabled people from exercising the same rights as abled people. However, as a type of closing remark, I think it is important to acknowledge some of the moral consequences of the use of PGD to select for any trait whatsoever.

There is a nascent trend that reveals what kind of shifts may occur in a population when technologies that allow for singular

trait selection are made readily available. In Denmark, the availability of pre-natal testing has drastically reduced the number of children born with Down syndrome, as 95% of mothers choose to abort fetuses that have been diagnosed with the condition. The result is a diminishing population of children with Down syndrome, and, consequently, a diminishing number of resources available to assist affected families (Zhang, 2020). While deaf families can be empowered by technology like PGD, and this can prove to broaden and strengthen diversity by allowing for more opportunities for flourishing in the deaf community, it also has the repercussion that people in general can more easily avoid giving birth to children with disabilities such as Down syndrome, achondroplasia, and deafness. This presents conflicting desires about two values most will acknowledge as vital, namely the importance of protecting reproductive autonomy for parents, and the active dismantling of an ableist master narrative. While to determine which of these values should take precedence is beyond the scope of this paper, I hope to leave the reader with some thoughts about the responsibility of parents who select for any trait whatsoever.

Dena Davis closes her argument against the use of PGD by asking us to consider the motivations behind a less controversial choice, like selecting for anatomical sex characteristics, in an embryo:

> Parents whose preferences are compelling enough for them to take active steps to control the outcome, must, logically, be committed to certain strong gender-role expectations. If they want a girl that badly, whether they are hoping for a Miss America or the next Catherine McKinnon, they are likely to make it difficult for the actual child to resist their expectations and to follow her own bent. (Davis, 1997)

It's possible to recognize a moral distinction in the argument of selecting for any preferred traits in children. One cannot escape the fact that there is a strong desire to fulfill some parental expec-

tation by making such a decision. One might ask if these sorts of decisions are made with the child's best interests in mind, or as a means of bringing some desire of the parent to fruition. In short: is such an action simply treating the future child as a means to an end?

One may counter this concern by bringing up the reality that it is essentially impossible for a parent to raise a child without any hope for personal fulfillment of some sort. It is also impossible to raise a child with full respect to their autonomy, and it is undoubtably true that, in many cases, it may even be harmful to the child to deprive them of certain forms of autonomy-limiting guidance—even if that guidance also serves the specific desires of the parents. Parents will inevitably want something out of the relationships they have with their children; it is possible to conclude that it is not entirely morally blameworthy for a parent to seek personal satisfaction from raising their child in a specific way.

However, as a response, it could be said that while one cannot fully avoid raising a child in a self-satisfying way, it's prudent to take into account how drastically one prioritizes one's personal satisfaction over a child's ability to self-govern. If it's possible to recognize that preserving a person's autonomy to a significant degree is a good thing, parents should also be prepared to take responsibility for any decisions made on behalf of their children that prioritize their own projected interests over the child's freedom.

Wherever the reader stands on the issue, I hope that I have made a convincing case that disabled voices are a necessary part of that conversation.

Notes

1, In her paper, Wallis writes, "In one of the few investigations stratifying parental preferences with respect to disability more generally, Baruch et al. (2008, p. 1055) report that, in the United States, '[t]hree percent of IVF–PGD clinics report having provided PGD to couples who seek to use PGD to select an embryo for the presence of a disability.' Note that this percentage

represents couples who sought to use PGD in this way, not the percentage of cases in which selection for disability actually occurred; nonetheless, this report suggests that the preference to select for disability is neither widespread nor trivial." This suggests that there are a significant number of deaf couples or individuals seeking to produce a deaf child through PGD.

2, Ableism is a prejudicial belief or attitude that asserts that abled people are in some way superior or more valuable than disabled people. This may be reflected through discriminatory practices, negative conceptions of disability, or many other types of harmful actions.

3, An anonymous referee wishes to address the idea that the inability to access certain aural pleasures, like music, should be considered one of these main arguments. Firstly, this assumes that one must be able to hear to enjoy music in any capacity, and does not take into account the vibrational and rhythmic qualities of music still available to deaf individuals. Secondly, a child who is born deaf has never been able to hear music, and thus cannot feel as if they have lost some important source of pleasure, having never been able to experience it in the first place. It is strange to think every person should be entitled to every single form of pleasure possible just by virtue of being born. Finally, simply because some pleasurable qualia are unavailable to certain people does not render those individuals' lives less valuable overall. For example, a person with a heart condition might not be able to ride roller coasters, and so that particular pleasure isn't open to him, but his life can still be very full with the other pleasures that *are* available to him. A person who is allergic to chocolate has a variety of other flavors to enjoy—one would not say that her life is somehow emptier or deprived because she does not have access to certain qualia that the majority of people enjoy. Because of this, it does not seem a particularly strong argument to think a deaf child's life is vastly less desirable simply because they cannot hear musical tones.

4, Davis discusses a 1972 court ruling in the *Wisconsin v. Yoder* case. In this situation, the court ruled to preserve the Amish way of life by exempting children from mandatory education past the 8th grade. This is evidence of the general attitude that many will want to honor the freedom for a parent to raise a child in an unusual but culturally important way. This is also evidence that our legal system generally supports the permissibility of such decisions.

5, An anonymous referee states that many people might personally disapprove of the lifestyles discussed here and consider their actions to be harmful to the children in question. I think this may be a misunderstanding of the principle of liberal neutrality; the core concept is that all reasonable interpretations of a good life should be allowed to be practiced without interference, even if they are not personally attractive to the majority. While many may consider it sad that Jehovah's Witnesses might not get to celebrate birthday parties as a child, it is still widely held that the parents' right to raise their children in a certain way should be upheld, even if it is not attractive to the majority.

Personal disapproval of a lifestyle does not mean that it should be considered morally impermissible, punishable, or illegal.

6, An anonymous referee asks, "How do we know these two embryos would actually be different people? What if a parent would give the child the same name in either scenario, couldn't either embryo become the same person, regardless?" This seems to me an error in reasoning; there is no pre-existing entity, that we definitively know of, that could somehow manifest as either embryo. The same name given to either embryo or the same parenting plan enacts on either does not, de facto, make them an identical being.

7, An anonymous referee brings up the fact that people cannot select for race in one's child, but I believe this question misses the point I was trying to make. The idea is not that a person could have somehow chosen a different race for their child, but rather that it is still considered morally permissible to choose to have a child under conditions of predictable societal disadvantage. This argument is used to show the view that, by the standards of liberal neutrality, it is reasonable to allow the birth of children who will experience disadvantage. Inevitable disadvantage should not be a deterring factor in choosing whether to bring a life into the world or not.

8, Again, this is not to say that we don't see societal disapproval for an individual who chooses to have a child in very difficult economic conditions, but rather that our laws and principle don't reflect that their decision should be completely impermissible or illegal.

9, Historically, eugenics was a process that aims to exclude certain genetic traits in human beings, with the goal of creating a homogenous group of human beings who display traits deemed desirable by the perpetrating social group. Eugenic ideology was infamously employed by the Nazis during World War II, and as a result millions of lives were destroyed in the name of curating a supposed racial and genetic "purity." Disabled people were counted among those who were targeted during this genocide.

10, Projective synesthesia is a perceptual condition that causes a person to experience an involuntary sensory association when stimulated. For example, when a certain word is spoken, the synesthete may experience a smell or taste that is associated with it. Some synesthetes can see music as shapes and colors, or experience tactile senses accompanied with certain auditory inputs.

11, I would like to thank Dr. Talia Bettcher for her assistance in helping me to develop many of the points I used in this paper.

Bibliography

Davis, Dena. (1997) "Genetic Dilemmas and the Child's Right to an Open Future," *Hastings Center Report* 27(2), pp. 7–15

Feinberg, Joel. (1980) “The Child's Right to an Open Future,” in William Aiken and Hugh LaFollette (Eds), *Whose Child? Children's Rights, Parental Authority, and State Power*, pp. 124-153 (Totowa, N.J.: Littlefield, Adams & Co.)

Häyry, M. (2004) “There Is a Difference Between Selecting a Deaf Embryo and Deafening a Hearing Child” *Journal of Medical Ethics* 5, pp. 510-512

Puga-Gonzalez, Cristian. (2019) “Liberal Neutrality and the Nonidentity Problem: The Right to Procreate Deaf Children,” *Journal of Social Philosophy* 20, pp. 363-381

Scully, J.L. (2008) *Disability Bioethics: Moral Bodies, Moral Difference* (New York, NY: Rowman & Littlefield)

Stramondo, Joseph. (2020) “Disability and the Damaging Master Narrative of an Open Future,” *Hastings Center Report* 10.1002/hast.1153, 50, S1, (S30-S36)

Wallis, J.M. (2020) “Is it Ever Morally Permissible to Select for Deafness in One’s Child?” *Med Health Care and Philos* 23, pp. 3-15

Zhang, Sarah. (2020) “The Last Children of Down Syndrome”, *The Atlantic*, Dec 2020

Unity of Virtue and Wisdom in the *Protagoras*

John Hurley

Introduction

The history of virtue ethics is most often traced back to Aristotle, especially his *Nicomachean Ethics*, but Aristotle's approach to virtue is deeply stamped with Plato's influence. Among Plato's key ethical works is the *Protagoras*. As in such other dialogues as the *Laches* and *Euthyphro*, Socrates[1] assumes that (a) virtue (ἀρετή) is both meaningful and desirable and (b) virtue as a whole includes a number of specific virtues, including but perhaps not limited to courage, justice, temperance, holiness[2], and wisdom, which I shall call the "enumerated virtues." I will argue in this paper that Socrates's attempts in the dialogue to find some commonality between the different virtues do not succeed, and that the text suggests that Plato did not mean the dialogue to offer a satisfying ethical doctrine.

The *Protagoras*, like many of Plato's other works, opens with a practical question embedded in a dramatic context; should an ambitious young man pay a well-known sophist to teach him ἀρετή? As is typical, the question quickly devolves to deeper problems of definition: is virtue a form of knowledge and thus teachable, and how are the enumerated virtues related to each other? Subsidiary issues include: do the virtues differ in kind from each other, or are they homogeneous parts of the whole? What does Socrates mean when he predicates the virtues of themselves and each other, apparently in an interchangeable way (for example, when he claims that justice is just (330c-d), holiness is holy (330d), justice is a holy thing, and holiness is a just thing

(331b))? Does Socrates succeed in showing that whoever has one virtue must have all of them (for example, that it is impossible for an unjust person to be temperate or a coward to be pious)? Are some virtues subsets of others? If all the virtues reduce to wisdom or to subsets of wisdom, why is wisdom separately enumerated as a virtue?

As usual in the dialogues, Socrates asks leading questions designed to draw his interlocutor into contradictions. He proposes the positive claims that the different virtues are unified in some way and that wisdom underlies all the other virtues. However, he suggests several inconsistent accounts of the unity of the virtues. In the context of socratic dialectic, it is not always clear which of Socrates's proposals represent his, or Plato's, actual positions and which are feints intended to mislead the interlocutor. The candidate views include:

1. Virtue is "one single thing, but justice and temperance and holiness [are] parts of it" (329c).
2. Justice, temperance, and holiness "are all names for one and the same thing" (also 329c), and "wisdom and temperance and courage and justice and holiness are five names for one single thing" (349b).
3. The different parts of virtue are like "parts of gold, which don't differ at all from one another and from the whole except in largeness and smallness" (329d).
4. "Justice is such as holiness and holiness such as justice" (330b).
5. Justice is "either the same as holiness or a thing most like it" (331b).

Socrates also proposes the following view, which he explicitly rejects:

6. The different parts of virtue are "like the parts of a face," i.e.,

diverse parts that make up a whole (329d), and "a certain peculiar nature and reality underlie each of these names, a thing in each case with its own power, one not being such as any other" (349b).

Protagoras agrees with Socrates that wisdom is the most important of the virtues, (330a) but this is fundamentally different from claiming that the other virtues reduce to wisdom. He waffles between the views that the different parts of virtue are essentially different from each other, standing in relation to virtue as a whole "as the parts of the face are to the whole face," (329e) and that courage is unlike the other virtues, which are unified in some way (349d). He agrees with Socrates's dialectical suggestions that each virtue has its own "power" different from the others and that no virtue is "such as" any other (330b). At 349d, he claims that four of the virtues are "fairly like each other," but "courage is much different from them all" as manifested by the fact that "many men are extremely unjust and unholy and intemperate and foolish, but surpassingly courageous."

The balance of this paper is organized as follows. In Section I, I discuss some important modern scholarly views of the unity of the virtues in the *Protagoras*. In Section II I examine why the claims of interpredicability and identity are so counterintuitive, especially to modern readers. In Section III, I discuss some weaknesses in Socrates's arguments that lead me to my view that Plato does not mean any of Socrates's accounts of unity to be conclusive. My conclusions in Section IV point to some broader implications of my thesis.

Section I: Contemporary Views of the Unity of Virtue in the *Protagoras*

A number of recent scholars have written on this issue. Many are motivated by a desire to read Socrates's comments in such a way as to find a solid Socratic or Platonic doctrine, but I am not trou-

bled by the idea of a Socrates, or a Plato, who is a proto-skeptic. It is not implausible that Plato intentionally constructed the dialogue in such a way that Socrates does not offer a view that is more satisfying than his opponent's. This certainly describes the Socrates of many early dialogues, and I do not see a good reason why we should be afraid to conclude that it also describes the Socrates of the *Protagoras*. On the other hand, even in aporetic dialogues which fail to find definitions of such virtues as piety, Socrates always shares the interlocutors' assumptions that the named virtues, whose natures we admittedly do not understand, are good things. We might be reluctant to attribute such question-begging to Socrates or to Plato, and accordingly assume that a doctrine is to be found that is at least sufficiently complete to support the idea that these virtues are desirable. However, it means little if a doctrine is not actually present. It is also plausible that Socrates changes his mind about the unity of the virtues during the course of the *Protagoras*; indeed, he explicitly changes his mind about whether virtue is teachable.

Some articles on this topic argue for particular interpretations of the *Protagoras* on the grounds that Socrates's ideas therein must be consistent with his views in the *Laches*, *Euthyphro*, and other dialogues. Denis O'Brien (2003) also mines the *Timaeus*, *Laws*, and *Sophist* for help in understanding the *Protagoras*. Daniel Deveraux (2006) further cites Aristotle's and Xenophon's reports in trying to find a unified Socratic doctrine (Deveraux 2006, p. 335), though he also cites later dialogues to show changes in Plato's ideas over time. Yet, Socrates or Plato would not be the last philosophers to take inconsistent positions at different times. Socrates might try out different ideas in different discussions, and Plato might change his mind between the dates of composition of different works, either about the correct interpretation of the views of the historical Socrates or about his own positions. Furthermore, I want to avoid arguments involving dialogues other than the Protagoras for the practical purpose of limiting the

scope of this paper.

Recent discussions typically credit Gregory Vlastos (1972) and Terry Penner (1971) for initiating this discussion. Vlastos categorizes and names Socrates's apparently incompatible proposed accounts of the unity of the virtues as follows. The enumerated virtues may be:

1. Five names for the same thing (the Identity Thesis)
2. Similar to each other and similar to virtue in general (the Similarity Thesis)
3. Necessarily coextensive in their instantiations (the Biconditionality Thesis, also called the Equivalence View by Brickhouse and Smith (1997) and others). Note that this thesis is entailed by the identity view, but it does not itself entail identity.

Socrates's proposals that the virtues are "five names for one single thing" (made most clearly at 349b) or like "parts of gold, which don't differ at all from one another and from the whole except in largeness and smallness" (329d) certainly entail interpredicability, but they appear to be assertions of the Identity Thesis. Vlastos argues that the Biconditionality Thesis is sufficient to explain all of Socrates's claims about unity. The five enumerated virtues are not necessarily identical, but they are interpredicable. According to Vlastos, a statement like "justice is pious" does not mean that justice and piety are the same thing, or that the universal "justice" has the characteristic of being pious. Rather, for any human being, if any one of the virtues can be predicated of her, so can all the others. Whoever is just is also pious, wise, temperate, and courageous. Vlastos refers to this as "Pauline predication," in reference to St. Paul's comment that "charity suffereth long and is kind."[4] Paul's meaning is that, where charity is instantiated in a charitable human being, this person is necessarily also long-suffering and kind. Vlastos' claim is that, for Socrates, the same

relationship holds between any pair of the enumerated virtues. Furthermore, the relationship is reflexive; whoever is courageous is wise and whoever is wise is courageous.These relations are necessarily true; as Vlastos puts it,

$$\Box (x) (Cx \Leftrightarrow Jx \Leftrightarrow Px \Leftrightarrow Tx \Leftrightarrow Wx)$$

where Cx, Jx, Px, Tx, and Wx mean that x is, respectively, courageous, just, pious, temperate, and wise. (Vlastos, p. 424) $\Box$ is a modal operator indicating necessity, so this makes a stronger claim than an otherwise identical formulation using a universal quantifier in place of the box. It is not just the case that the set of pious people is, contingently, identical to the set of just people, it is also necessary that every pious person must be just and every just person pious. Interpredicability is true because the possession of wisdom is both necessary and sufficient for possession of all the other virtues.

Penner argues for the stronger identity thesis:

> Bravery = wisdom = temperance = justice = piety
> (Penner 1971, p. 79)

On Penner's view, Socrates is investigating neither the meanings of the names of virtues nor the essence of virtue, but the psychological state that leads people to virtuous acts. The question "what is courage," for example, is a "general's question," aimed at identifying why some people are courageous, not a "philosopher's question" aimed at a metaphysical or linguistic understanding (Penner 1971, pp. 81-83). A person with practical wisdom of good and evil would undertake actions that were courageous, just, etc., according to the circumstances (ibid, p. 88). More recently, Justin Clark (2018) offered a related view; the question is about the "psychological processes, mechanisms, and capacities involved in virtuous action" (Clark, 2018 p. 1325). Virtue is "a capacity (δύναμις) within the psyche of the virtuous person, a capacity of correct choice" (ibid, p. 1330).

Denis O'Brien and Allison Piñeros Glasscock (2020) observe that, under the conventions of Socratic dialectic, Socrates does not need to choose any particular account of the unity of the virtues, only to catch Protagoras in a contradiction. I believe that Socrates does not, in fact, offer an account of unity that we are expected to endorse.

Brickhouse and Smith argue that, under what they call the Equivalence Thesis (i.e., biconditionality), wisdom as an entity distinct from the four types of wisdom associated with the other four virtues is irrelevant. (Brickhouse and Smith 1997, p. 315) However, Socrates does not claim that the enumerated virtues comprise all virtues. It could plausibly be claimed that the inclusion of wisdom as a separate category makes it easier for this formulation to extend to additional virtues.

Section II: The Counterintuitiveness of Interpredictability

A thesis that is easily accepted by the reader hardly requires Socrates to defend it, and the claim that the virtues are interpredicable was presumably intended to be paradoxical. This section will examine some reasons why interpredicability is so difficult to accept. Some of the discussion here reflects distinctly modern concerns, but Protagoras himself argues, or at least asserts, the commonsense claim that some people are courageous but lack some of the other virtues (329e).

It is important in the discussion below that, if the enumerated virtues are species of wisdom, correct intentions are not sufficient conditions for virtue. Someone who intends to be courageous but has mistaken judgments about what to fear, or someone who intends to be just but has a mistaken view of justice, lacks these virtues. Correct intentions, though, are probably necessary for virtue-as-wisdom—there is no obvious account on which someone can be unintentionally wise.

If the enumerated virtues other than wisdom are species of wisdom and, in addition, interpredicability is correct, anyone who is wise in the way required by any virtue is wise in four apparently diverse ways. This does not amount to a claim that wisdom is monolithic, or even that the specific types of wisdom involved in the enumerated virtues make up a monolithic whole; it is consistent with the possibility that these different types of wisdom have different intensions but the same extension.[5] However, it does imply that a virtuous person is globally wise and an unvirtuous one globally unwise.

In addition to the problems posed by the difficult view of wisdom pointed to above, a different historical context shows the enumerated virtues in a different light. Courage is much less important to the typical modern reader of philosophy than it was to Plato's original audience of upper-class men who spent much of their time in military service. We may even be inclined to say that, while the concept "virtue" implies that a quality is a good one, the world would in fact be better off if military courage was more scarce than it actually is. Further, on a naturalist view holiness is problematic as a virtue, but I would not want to be hasty to conclude that the other virtues are irrelevant.

Using ordinary meanings of the modern English words "just," "pious", "temperate", "courageous," and "wise", and accepting Socrates's definitions of the first four as species of the last, we can easily find examples of people who apparently show some of these virtues but not others. It is also not unusual for people to exhibit particular virtues intermittently or inconsistently. We encounter plenty of people whom we would regard, on ordinary meanings of the virtue terms, as temperate but unjust, and others whom we might view as just but intemperate. It is even easy to think of examples in which we would judge people to show one virtue while, in the same act or pattern of behavior, manifestly lacking one or more of the other enumerated virtues. Any of us can easily name a war in which soldiers served with apparent courage in a

cause we regard as clearly unjust. The concept "unjust war effort" may not have been intelligible to Socrates or Plato, but it is not anachronistic to consider how interpredicability or unity apply to a case which we would view in that way.

It also appears possible, in commonsense terms, to be temperate in injustice. In my business career, I witnessed many cases in which colleagues exploited information asymmetries to the detriment of consumers, but I doubt most of those people would burglarize their customers' residences even if they could do so with impunity. Most soldiers refrain from committing atrocities even while serving in apparently unjust causes.

Temperance in particular is prone to internal inconsistency: a person who is temperate in one area may be intemperate in another. Some people eat healthy diets but use psychoactive drugs in excess; others are prone to excessive emotional outbursts but shun alcohol.

Holiness, depending as it does on specific beliefs about supernatural matters, poses a sharp issue for interpredicability. According to Socrates, holiness is either a subset of wisdom or identical to the whole of wisdom. But suppose believer A and believer B subscribe to religions with contradictory doctrinal claims. Believer A might think that believer B is unwisely sacrificing to gods that do not exist, and perhaps even committing sacrilege in doing so. Believer B might think believer A is following unnecessary behavioral rules based on mistaken beliefs. Each would claim that the other, lacking wisdom in the matter of the divine, is failing in his attempts at holiness. If holiness is wisdom about dealing with the gods, then at most one of the believers is holy. If at least one of the believers is correct about the other, at least one is not holy, and if they are both incorrect, they both fail to grasp the type of wisdom that constitutes holiness and are thus both unholy. A and B, however, might both be uncontroversially temperate, courageous, and just. They certainly might also both be wise in matters not covered by the virtues; they might, for

example, both be skilled sculptors.

A naturalist would also want to separate holiness from the rest of the enumerated virtues on the more radical grounds that it is simply an empty exercise. A truly wise person, on a naturalist account, might in some circumstances choose to appear holy, but would neither actually be nor even attempt to be holy. Worse, if holiness is a form of wisdom and there are no gods, then no one actually *could* be holy. The fact that neither Plato nor Socrates adheres to a naturalist view excuses them from a charge of inconsistency, but it raises a more important issue: if one of the enumerated virtues is impossible, interpredicability entails that no one has any of the other virtues either.

Since, using normal meanings of the words used in modern English to describe the enumerated virtues and viewing the virtues as species of wisdom, we can find examples that appear to show interpredicability to be false, at least one of the following must be true:

1. Some of our words "holiness", "justice", "temperance", "courage", and "wisdom" do not convey the intended meanings of Plato's Greek words. This might be due to either of the following two issues:
    a. The modern English words listed above are incorrect translations.
    b. Socrates or Plato has in mind non-obvious definitions of these words.
2. Our commonsense judgments are mistaken. Specifically at least one of the following is often true:
    a. People lack specific virtues, but appear to have them.
    b. People have specific virtues but appear to lack them.
3. Interpredicability is false. This implies one of the following:
    a. Socrates is not committed to interpredicability or to the

stronger claim of identity, which entails interpredicability, or

b. Socrates is committed to interpredicability, and perhaps to identity, but his arguments are incorrect.

Alternatives 1 and 2 could help Socrates's accounts of the unity of the virtues, but would not resolve the inconsistencies in his arguments I will discuss later in this paper. I will argue below that Plato has constructed the dialogue in such a way as to suggest that either alternative 3a or 3b is correct.

Section III. Socrates, Plato, and The Unity of Virtue

There are several senses in which a dialogue may be aporetic. It may be weakly aporetic in the sense that, in the dramatic context, the dialectic does not lead to a resolution that all parties support; in this case, an account may nonetheless be offered which is intended to be convincing to the reader. A dialogue may be aporetic in the stronger sense that no account is offered that Plato intends the reader to find fully satisfying. A dialogue may also be very strongly aporetic in the sense that no coherent account is offered at all. I will argue that the *Protagoras* is aporetic in the second of these three senses.

Socrates offers several accounts of varying coherence, but whether he is committed to biconditionality, identity, or some other account of the unity of the virtues, Plato does not intend Socrates's arguments to settle the issue. This is shown in several cases in which Socrates either contradicts his own arguments or calls attention to their weaknesses in more oblique ways.

Socrates undermines interpredicability by giving several commonsensical expositions of disunity among arts and professions while presenting these as forms of knowledge. At 311b-312b, he asks his friend Hippocrates what he hopes to learn from a course of study with Protagoras. Socrates questions Hippocrates

about two professions he might learn from well-known specialists, namely medicine and sculpture. The character Hippocrates is not the famous physician, but the fact that Plato has gone so far as to give him that name suggests that he wants to signal the importance of this issue. Socrates then observes that his friend Hippocrates has also learned other disciplines, namely writing, music, and physical training, from teachers, though without aspiring to become "a public practitioner" of these arts. These five fields are presented as fields of knowledge, and therefore subjects that can be taught, in a setup for the discussion about whether Protagoras will be able to teach virtue. Skills in sculpture, medicine, writing, and music are obviously not interpredicable; otherwise, there would be no point in seeking out practitioners of medicine or sculpture to study their arts, since a good music teacher would be sufficient.

Socrates continues in this vein at 312d, where he asks Hippocrates which kinds of wisdom sophists might be able to teach. He says that "if someone asked us what wise things painters know, we'd surely reply that they know about the production of likenesses, and so in other cases." Socrates returns to this theme in his soon-abandoned argument that virtue is not teachable; the Athenians, he says, call on specialists when their arts are relevant to public debate, ignoring others who may have higher status but no expertise, but they assume that everyone is qualified to speak on matters of virtue, rather than seeking out specialists in the field (319b-d). Socrates will change his mind about the teachability of virtue, but this does not undermine the relevance of the observation about arts and professions.

If categories of wisdom are not interpredicable, the claim that wisdom is interpredicable with the other virtues requires modification. A subset of wisdom consisting of those aspects that are involved in the virtues could be substituted for wisdom in general. It could be posited, for example, that a holy person necessarily possesses the wisdom involved in the other virtues and that someone who possesses such wisdom is necessarily holy.

Alternately, it could be allowed that, while wisdom is not monolithic, interpredicability involves only people who are globally wise. For example, people who possess all types of wisdom would be necessarily temperate, and temperate people would necessarily possess global wisdom, but some people might lack temperance yet possess certain kinds of wisdom; an intemperate person might nonetheless be an excellent musician. However, Socrates does not explore either of these possibilities in the *Protagoras*.

Since Socrates contradicts interpredicability, he also contradicts the stronger identity view, which entails interpredicability. His arguments are consistent with some weaker form of similarity between the virtues, but, although he uses such locutions as "more probably justice is either the same as holiness or a thing most like it, but it is beyond question that justice is such as holiness and holiness such as justice;" (331b) and "justice and holiness are pretty nearly the same thing," (333b) he never proposes a thorough account of such weaker types of similarity.

In his examination of Protagoras, Socrates gives the first hint of his argument for the unity of virtue at 331b, where he raises the specter of a critic who asks whether, if the virtues are different from each other, "holiness is not such as to be a just thing, nor justice such as to be a holy thing, but such as not to be a holy thing? And holiness such as not to be a just thing, but perhaps then an unjust thing, [that][7] is unholy?" Socrates soon launches a line of questioning at 332a-333a that is intended to establish his argument that the virtues are unified because they share the same opposite. Wisdom is the opposite of folly, but people who act intemperately act out of folly. But "there is only one opposite... for each one of the opposites, and not many?" Therefore, wisdom and temperance are identical.

The argument from opposites isn't sound because the claim that every opposite has only one opposite is incorrect. An entity may have multiple opposites that we identify under different circumstances depending on the granularity with which we examine the

opposition. If temperance is a proper subset of wisdom, we may view either folly or temperance, a proper subset of folly, as its opposite. It could be that temperance and piety are both parts of wisdom, while intemperance and impiety are both parts of folly. At a high level of granularity, the opposite of temperance is intemperance and the opposite of piety is impiety. On a less granular view, folly is the opposite of both temperance and piety. To save the argument as it applies to temperance and wisdom, we need the additional premise that temperance is the whole of wisdom. To make sense of the argument as it applies to temperance and piety together, we need to assume that intemperance and piety are not two different proper subsets of folly. Either of these may be correct, but both beg the question.

Socrates himself contradicts the argument from opposites at 341a-e. This occurs during the lighthearted discussion of Simonides, but what he says here sheds light on the serious argument about interpredicability. Socrates says that Prodicus criticises him for using idioms like "terribly wise" on the grounds that "what is terrible is bad," and that he should be "embarrassed to call good things terrible." For this reason, he says, "no one ever says 'terrible wealth,' or 'terrible peace,' or 'terrible health.'" So far, we have three entities that are "good" and not "terrible," but he then compares these usages to a claim that "it is bad to be good," implied by Prodicus's (unserious) gloss of "hard" as "good." Socrates himself, then, implies that "terrible" is not just opposite to "good," but also to "wise," and perhaps to "wealthy," "peaceful," and "healthy." This points to the scope ambiguity in the argument from opposites itself.

Socrates contradicts the argument from opposites again and more directly at 349e-350c, in the discussion of wise and foolish confidence. "Some people dive boldly into wells," confident "because they know how." Some are "bold and confident in making war from horseback"

and some are ready to "fight with light shields," a skill

glossed as "skirmishing." While generally people who are skilled in these various things are more confident than those who are not, Socrates and Protagoras agree, there are some people who are confident in these matters despite lacking the necessary skills, and such people are not courageous but mad. The skills of well-diving, cavalry fighting, and skirmishing are certainly not all names for the same thing. Nor are incompetent well-diving, incompetent cavalry fighting, and incompetent skirmishing, but Socrates presents these as the multiple opposites of "madness," again pointing to the scope ambiguity in his argument from opposites.

At 353c-355c, Socrates makes a hedonistic argument that those who do evil do so from ignorance, rather than being overcome by pleasures. Pleasure and pain neutralize each other; some things that are immediately pleasurable may be bad because they result in greater net future pains, and some things that are immediately painful may be good because they result in greater net future pleasures (353c-354e and 356a-c). Socrates induces Protagoras to agree that "pleasure is good and pain is evil" (354c) and later states that "good" and "pleasure" are synonymous (355c). Since no one would voluntarily choose a course of action that results in net pain or loss of net pleasure, those who do evil when given a choice must do so out of ignorance. This argument will figure in another argument from opposites.

Socrates sets up the new argument at 358d-360e, where he attempts to show that no one can be both foolish and courageous. No one "goes towards what he believes fearful," but the courageous go to war because they are aware of the net future benefits of doing so, while cowards refuse to fight because they are ignorant of those benefits (359c-360a). Cowardice is "ignorance of what is and is not to be feared" (360c). Opposite to this is courage, as well as "wisdom of what is and is not to be feared," (360c-d) so these two must be identical. He then asks Protagoras whether he stands by his claim that some people are both foolish and courageous. Socrates has manifestly not shown here that no courageous person

can be foolish about anything, only that they are not foolish about what is or is not to be feared.

Conclusion

I have argued that, in the *Protagoras*, Socrates undermines his own arguments for interpredicability of the five enumerated virtues and thus for the stronger claim that the virtues are identical, and more broadly that the dialogue is not intended to convince us that interpredicablity is correct. It is possible that Socrates is not committed to all his arguments, and that some are feints intended to show inconsistencies in Protagoras' claims. It is also possible that Plato means us to think that Socrates is committed to these positions, but that his arguments are unsuccessful. In my view it is much more likely that the weaknesses are intended by Plato than that such a consistently subtle author is inattentive to the weaknesses of the arguments he puts in his mentor's mouth. The issue remains unsettled not just for the dialectical opponents, but for Plato and for us.

Ethical theories generally attempt to build from statements that purport to summarize the essence of goodness. The basic principles of Kantian and Utilitarian ethics, for example, can each be coherently summarized in one or a few sentences that provide bases for their more specific and developed claims. Virtue ethics might be summarized by a formulation similar to this: "ethics consists of acting in accordance with virtues such as ...", with a list of virtues appended. Such a formulation, however, lacks coherence as a definition unless some common essence can be found for the various virtues; there must be something all virtues have in common that justifies treating them all as virtue. Even if separate virtues are unified simply in that they are all good characteristics for human beings to have, some common criterion for goodness is needed. The *Protagoras* offers an inconclusive preliminary inquiry, not an embryonic theory of virtue ethics.

Notes

1. It is, of course, unclear how closely Plato's characters Socrates and Protagoras reflect the views of the historical persons Socrates and Protagoras.
2. In quotations from the secondary sources and from translations of the *Protagoras* itself, I shall follow the sources' differing usages of the words "piety" and "holiness" as translations of the Greek words εὐσέβεια and ὁσιοτής.
3. I use R.E. Allen's 2008 translation of the *Protagoras* throughout this paper.
4. 1 Corinthians 13:4-7, King James translation.
5. Brickhouse and Smith point this out (pp. 316-317), although they argue on the basis of passages from the *Euthyphro* and the *Laches* that it is not Socrates's position. It is certainly compatible, if not synonymous, with Vlastos's view.
6. As a beginner at Greek, I shall have little to say about this possibility.
7. The bracketed "that" replaces the word "what" in Allen's text. I believe the "what" to be a typographical error.

Bibliography

Brickhouse, Thomas C., and Smith, Nicholas C. (1997) "Socrates and the Unity of the Virtues," *The Journal of Ethics*, 1:4, pp. 311-324

Clark, Justin C. (2018). "Socratic Inquiry and the 'What-is-F' Question," *European Journal of Philosophy* 26, p. 1324-1342

Deveraux, Daniel. (2006) "The Unity of the Virtues," in: Hugh Benton (Ed), *A Companion to Plato*, pp. 325-340 (Hoboken, NJ: Wiley-Blackwell)

Glasscock, Allison Piñeros. (2020) "The Discipline of Virtue: Knowledge and the Unity of the Virtues in the *Protagoras*," *Ancient Philosophy* 40, pp. 41-65

O'Brien, Denis. (2003) "Socrates and Protagoras on Virtue," *Oxford Studies in Ancient Philosophy* 24, pp. 59-131

Penner, Terry. (1971) "The Unity of Virtue," *Philosophical Review*, 80. Reprinted in Gail Fine, (Ed) (1999), *Plato 2: Ethics, Politics, Religion, and the Soul*, pp 78-104 (Oxford: Oxford University Press)

Plato. *Euthyphro*, *Laches* and *Protagoras*. (2008), R.E. Allen (TR), in: *The Dialogues of Plato*, vol. 3 (New Haven: Yale University Press)

Vlastos, Gregory. (1972) "The Unity of the Virtues in the *Protagoras*," *The Review of Metaphysics*, 25:3, pp. 415-458.

Human Rights, Universality, and Moral Desert

David Randall

Introduction

Since the Universal Declaration of Human Rights (UDHR) was proclaimed by the U.N. General Assembly in 1948, human rights (HR) have played an increasingly important role in international politics. HR violations are often used to justify the intervention of one state into the affairs of another. This intervention can be military, diplomatic, or economic (sanctions, for example). In addition, HR law creates an international legal framework to hold states accountable for their behavior, and compliance with these laws can serve as a precondition to political legitimacy. As the role of HR continues to grow, controversies over which rights qualify as HR have become common. Theories of HR attempt to mediate these controversies by proposing various foundations or justifications that can give rise to HR. One such justification proposed by Jason Nickel (2015) is moral desert; whether or not someone is morally deserving. Nickel claims that the notion of moral desert justifies some HR (giving rise to them or shaping their content) and qualifies some others (putting limits on who has those rights and the extent to which those rights can be limited) (Nickel 2015, pp. 157-162). Zofia Stemplowska (2015) disagrees, and offers convincing counter-arguments as to why HR cannot be justified or qualified by desert in most cases. However, more can be said about Nickel's arguments regarding HR whose contents include proportionality.[1] While his approach is intuitive, it needlessly threatens the universality of HR, among other things. I will attempt to show that a theory of HR need not appeal to moral

desert when proportionality is present, as Nickel claims it does, and that this can be done consistently with a broad number of possible foundations for HR while preserving many of HR's most appealing features.

Section I: The Question of Foundations and its Importance

Among the least controversial definitions of HR is that they are rights we have by virtue of our humanity. It follows from this definition that HR are universal; all human beings have HR simply because they are human. In addition to universality, some theorists of HR believe that they are inalienable (we cannot lose them), essential (they protect those interests most important to us as human beings), and that they hold a special place in the hierarchy of rights. In other words, HR may have more moral force than many other kinds of rights. It is not my intention to argue for these features here, although I support them. However, a good theory of HR would need to preserve whichever of these features its authors are committed to. Nickel's nomination of moral desert as a possible justification for HR undermines all of them.

As the language of HR has become more prominent in international affairs, more seems to be riding on how we can differentiate between HR and other kinds of rights. Many groups have tried to see their interests enshrined as HR.[2] States, on the other hand, often wish to keep the number of HR to a minimum for two reasons. First, states shoulder the burden of the duties that correspond to our HR, and those duties can be taxing. Second, HR can functionally limit the sovereignty of states by setting boundaries on their behavior; when violations occur, members of the international community are justified in intervening. Moreover, while some HR theorists might wish to expand the catalog of HR to include more protections, others worry that inflating the number of HR will devalue them, making them less effective.

One way to distinguish between HR and other rights is to identify HR's foundation or justifications and see whether any given right can sensibly be derived from them. Insofar as HR are those rights we have by virtue of our humanity, it is reasonable to assume that some feature of our humanity or human experience might give rise to HR. Proposals along these lines, such as human agency (Griffin 2008), basic human needs (Miller 2012), conditions for a good life (Liao 2015), human dignity (Waldron 2015), or human capabilities (Nussbaum 2011), are examples of possible foundations from which many HR can be derived. Nickel claims that some HR can be justified by moral desert (Nickel 2015, p. 160). I take him to mean that he is not a foundationalist; instead of positing one broad foundation undergirding all HR, he is implying that HR can be derived from various justifications. In his eyes, moral desert is one such justification.

Although the idea of HR is relatively young, the question of foundations and justifications cannot be settled by appealing to the historical record. The authors of the first canonical HR document, the UDHR, and the two covenants that followed in 1966, the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social, and Cultural Rights (ICESCR) were purposefully vague on this issue. They were motivated to generate broad international support for the HR movement, and putting forward a single philosophical foundation would have invited controversy between the fledgling U.N.'s member nations. Ultimately, they couched many of the articles in terms of human dignity, a concept sufficiently ambiguous[3] to satisfy everyone, and so the matter, with everything that follows from its resolution, is still open for debate (Luban 2015, pp. 275-276). That being said, HR practice is robust and its many benefits have been firmly established over decades of implementation. Therefore, another requirement of a successful theory of HR would be to preserve the majority of those rights we currently enjoy, as well as all of the rights that we consider to be paradig-

matic cases, such as the right not to be tortured or the right to political participation. Nickel proposes that moral desert preserves some of the HR laid out in the UDHR, ICCPR, and ICESCR.

I agree with Stemplowska that carving out a role for considerations of moral desert to play in a theory of HR would make such a theory rather less attractive than it needs to be by undermining important features of HR, including their universality. I also believe that there are compelling alternatives to accept considerations of moral desert that are capable of filling the role Nickel ascribes to them. These alternatives are compatible with a broad range of possible foundational approaches, making them more flexible than Nickel's proposal. In Section II, I will outline Nickel's arguments for why considerations of moral desert should play a role in a theory of HR, and provide examples in which he believes moral desert qualifies or justifies HR. In Section III, I will lay out Stemplowska's arguments against considerations of moral desert in a theory of HR, specifically those rights that do not include proportionality as a part of their content. In Section IV, I will address those rights that do include some measure of proportionality as a part of their content. First, I will show how the HR to proportional pay might be derived from the HR not to be exploited, which can follow from many if not all of the most promising foundational approaches. Second, I will argue that the HR not to suffer punishment disproportionate to a crime can be justified by forward-looking considerations that represent the HR of the victims, potential victims, and the offender themselves, as opposed to backward-looking desert considerations. In this case, proportionality is best understood as a way to balance competing rights or HR claims. In Section V, I will respond to some potential objections to my arguments, and in Section VI I will voice concerns I have about HR enforcement when moral desert comes into play.

Section II: Moral Desert as a Justification for or Qualification of HR

Nickel recognizes that qualifying or justifying HR with considerations of moral desert necessarily diminishes their universality by making possession of some HR contingent upon one's desert basis. However, he thinks that as we move from abstract features of HR to specific rights this becomes unavoidable (Nickel 2015, p. 172). He gives two arguments for why this kind of qualification or justification is permissible.

Nickel's first argument might be thought of as the existing qualifications argument. According to Nickel, HR are already qualified by a variety of considerations, including need, vulnerability, consent, and ability. For example, we would not say that someone in perfect health has the right to medical treatment as that treatment would be superfluous and unnecessarily taxing on the health care system (need), or that someone who has not been charged with a crime has the right to a fair trial since they are not at risk of wrongful conviction (vulnerability). Nor would we insist that someone's right to food could not be waived if they were carrying out a hunger strike (consent), or that children should have the right to serve in public office since they are not capable of doing so (ability). In light of this, moral desert is just one of many qualifications limiting the universality of HR, as opposed to the sole malefactor, and the relative strain it puts on HR's universality is negligible (ibid. p. 174). I will return to this in Section III.

Nickel's second argument could be called the relative equality argument. According to Nickel, nearly everyone's desert bases are practically equal where HR are concerned. We might imagine a bell curve, with some outliers on one tail of the curve having an exemplary desert basis. On the other tail of the curve, some small percentage of people have deplorable desert bases. Only these two groups would have their HR qualified by considerations of desert—the rest of us would fall in the center of the

curve and be unaffected. It follows from this, Nickel argues, that HR's universality is still largely intact; there are only some small exceptions to the rule (ibid, p. 179). If he is correct, relatively little harm would come from this kind of qualification. However, his argument does not justify that harm. In addition, it is doubtful that if the particular HR Nickel refers to are truly justified or qualified by moral desert that the tails would include a trivial number of people. To give just one example, Nickel claims that HR to participate in one's political system by holding public office is qualified by desert. In the United States, as of 2010 about 1 in 12 of all adults have a felony conviction preventing them from exercising this right (Shannon, et al. 2017). Twenty-six million people hardly amount to a trivial tail.

Nickel discusses at least five other HR (all of which are enshrined in legal and canonical HR documents) that he claims are qualified or justified by desert. There is little I could add to Stemplowska's objections regarding considerations of desert impacting the right to hold public office (ICCPR Article 25), the right to freedom of movement (ICCPR Article 12), or the right to due process (ICCPR Article 14) (Stemplowska 2015, pp. 168-172). I will say something about her approach generally in Section III. However, the right to be protected from punishments disproportionate to one's crimes (ICCPR Article 6.3, European Charter of Fundamental Rights in the Treaty of Lisbon 49:3) and the right to just and favorable remuneration for work (ICESCR 7, 15.1(c)) include some notion of proportionality, which must be justified somehow.

On the face of things, it might seem intuitive to say that criminal offenders should get the punishments they deserve and that workers should get the pay they deserve. I will not challenge Nickel's interpretation of the text of these articles here, although I think there is space to do so. Nor will I challenge whether or not these rights should be considered HR. Instead, I will address Nickel's position on his terms. I will argue that the apparent

proportionality inherent in these rights can be justified in ways that do not limit the universality, essential nature, or moral force of HR in Section IV.

Section III: Stemplowska's Arguments Against Considerations of Moral Desert in HR

Stemplowska's objection to the existing qualifications argument goes as follows. It may be the case that we qualify some HR on the basis of need, consent, vulnerability, or ability. However, the first three bases do not threaten the universality of HR in the same way that moral desert does. This is because no one loses the protections the rights were meant to provide (ibid. pp. 171-172). Consider need; the right to healthcare does not entail that a healthy person has a right to a heart transplant. However, should they experience heart failure, they do have a right to a heart transplant. Moreover, we all hold this right; its universality is not threatened by being qualified. Regarding consent, choosing to go on a hunger strike does not negate our right to food should we demand it. Temporarily waiving the right does not abrogate it—the protection remains. Vulnerability is similar. It would be absurd for me to demand a fair trial if I have not been convicted of a crime. That being said, it should be clear that my due process rights are still protected and that we all hold this right equally; the qualification does not limit the universality of the right.

Ability is a special case, and Stemplowska gives a different account (ibid, p. 172). She is a proponent of the interest theory of rights, a position articulated by John Tasioulas (2015). According to the interest theory, our essential human interests give rise to our HR. Instead of using this framework, I am going to make the argument in terms of rights in conflict in order to accommodate a broader range of possible foundations. This has no bearing on the content of the argument. Sometimes certain rights are qualified

on the basis of ability. For example, children and some mentally disabled people may not have certain rights related to their self-determination. In these cases, these individual's rights conflict with rights held by themselves or others. To accommodate the view that HR have stronger moral force than some other kinds of rights, we could even say that in these cases, the HR of the individuals in question are in conflict with other HR that they or others hold, such as rights whose objects are the safety and security of persons.

Section IV: Alternatives to Considerations of Moral Desert in HR with Proportional Content

Nickel maintains that whenever proportionality is a part of the content of a HR we should look to considerations of moral desert as potential justifications (Nickel 2015, p. 156)—because, surely, what we get should be proportional to what we deserve. In this section, I will argue against this position by addressing the right to proportional pay and the right against disproportionate punishment.

Section 4.1: The Right to Proportional Pay

Nickel claims that when we work, our desert status changes. In short, we have a right to be compensated for our labor, and that compensation is proportional to the desert generated through the contribution we have made (ibid p. 162). He points to Article 23(3) of the UDHR as expounding this right. It states, "Everyone who works has the right to just and favourable remuneration ensuring for himself and his family an existence worthy of human dignity, and supplemented, if necessary, by other means of social protection." Article 7 of the ICESCR employs the same language. To begin, it is not clear to me that "just and favourable remuneration" is the same as proportional remuneration[4], or what the remu-

neration is proportional to. But, putting the first concern aside, I believe that something besides moral desert can account for this instance of proportionality.

Stemplowska argues convincingly that the right to payment itself is generated by entering into a contract, independently of desert (Stemplowska 2015, p. 173). If I promise to pay you for mowing my lawn, you are entitled to that money regardless of whether you deserve it.[5] Ostensibly, if I employed you to do something legal, yet slightly odious, you would still have a right to be paid. Obviously, that right does not hinge on moral desert, as your actions in this case negatively impact your desert basis. However, once the right to pay has been established, the question remains as to what is accounting for the right to *proportional* pay.

Whereas Nickel contends that our compensation should be proportional to the desert generated through work, I would argue that the proportionality content of the articles above is derived from the right not to be exploited. Some HR theorists, including Liao (2015), Waldron (2015), and Gould (2015), have given accounts of how rights might reasonably be derived from one another. Cruft, Renzo, and Liao call those rights from which other rights are derived basal rights (Cruft, et al. 2015, p. 8). One way to think about what a basal right could be (among other things) is to consider the argument against some foundations of HR—that HR must be timeless, applying to human beings by virtue of their humanity independently of their circumstances or when they live. This makes some foundations less attractive since they might have more difficulty explaining rights that are context-dependent and exclude our distant ancestors, for example.

A remedy for this kind of attack is the notion of a basal right based on the foundation, from which a secondary, context-dependent right can be derived (I am not using secondary rights in the same sense that it is often used in HR literature, where a secondary HR would be a right whose corresponding duty is born by any entity other than a state). For example, Griffith's founda-

tion of autonomous agency might generate a basal right to self-realization, from which the secondary right to hold public office might reasonably be generated. To be clear, I mean nothing more or less by derivative than that these secondary rights could not have come to be considered HR without their basal counterparts. Another way to say this would be that violations of the secondary right are only HR violations if they also violate the basal righs.[6] Using this model, it could be said that the right to proportional payment is a positive secondary right, generated from the negative basal right not to be exploited. The right not to be exploited is itself consistent with and can be generated from a wide range of philosophical foundations. To name a few, exploitation is antagonistic to our autonomous agency, it can impinge upon basic human needs or the conditions for a good life, and it undermines human dignity.

The first benefit of seeing proportional payment as secondary to the right not to be exploited is that it preserves the essential nature of the right. If Jeff Bezos hired Elon Musk to mow his lawn and paid him half of the money he deserved for his labor, my intuition is that Musk's HR have not been violated. The kind of unfairness evident in this case is not essential enough, i.e., it does not protect an interest essential to Bezos as a human being, to generate a duty at the HR level. This seems to be supported by the text of the articles; “just and favourable remuneration” is meant to ensure “for himself and his family an existence worthy of human dignity.” To put it more broadly, the end goal is an existence that satisfies the moral considerations ascribable to the foundation or justifications the right is based in, whatever they may be.

The second benefit of this approach is that it opens up the question of what proportional pay can be proportional to. Under Nickel's account, the right is justified by desert and pay is therefore proportional to desert. If exploitation is underlying proportionality, then proportionality can track any criteria that, if not reflected in remuneration, might make work exploitative. These

criteria might include contribution, effort, agent responsibility, industry standards, the living wage in the locale the rights bearer inhabits, or any combination of these elements, to name a few. If desert underlies proportionality, it would still have to be "translated" from a criterion like those in the list above, in the sense that considerations from desert would have to flow from work in some way and the amount to be paid would have to flow from the amount of desert generated. It is not clear to me how this should be done or why it is desirable. After all, how is desert reliably ascertained and how much remuneration should it give rise to? More importantly, who should make these judgments? These questions hint at the ambiguity in the notion of desert itself. Moral desert as a theoretical concept does not provide any real criterion for answering questions like these, and desert judgments are highly subjective. Any systematic enforcement of a HR formulated as such would yield inaccurate and inconsistent results.

Section 4.2: The Right to Proportional Punishment

Article 6.2 of the ICCPR states that "in countries which have not abolished the death penalty, sentence of death may be imposed only for the most serious crimes...." and Article 49.3 of the European Charter of Fundamental Rights (ECFR) reads, "The severity of penalties must not be disproportionate to the criminal offence."From these articles, it is reasonable to assume that we have the HR not to face punishments disproportionate to crimes we might commit. What accounts for this proportionality? Nickel says, "Almost as bad as punishing the innocent is punishing with very severe criminal punishments a person who has committed a minor crime... Human rights requiring that punishments be proportional to crimes committed seem to be at least partially desert-based." (Nickel 2015, p. 160)

It may seem natural to assume that this is so—after all, what could account for the scope of a punishment besides the amount of

punishment an individual deserves? I will argue here that rights—or even HR—in conflict play that role more effectively than desert does.

Before I begin, Stemplowska reasons (correctly, in my estimation) that desert does not have to play a role in whether someone should be punished. Other considerations might be doing the work. I will outline her arguments briefly, focusing on incarceration, an example that both she and Nickel employ (Stemplowska 2015, pp. 170-171). First, some HR, including the right to freedom of movement and the right to self-determination, are impinged upon when their holders are incarcerated. Desert is not a necessary consideration for losing these rights (although it may or may not be a sufficient reason), as some mentally disabled people and children are not undeserving but lose them nonetheless. Second, forward-looking considerations, such as the protection or security of others or fairness to victims might be justifying punishment, as opposed to backward-looking considerations of desert. Third, if we understand moral desert as being a status that changes as we author moral acts, then there is no reason someone could not accumulate positive desert that would offset the negative desert generated by perpetrating a crime and therefore avoid prison. These arguments are compelling, but once punishment has been adequately justified, it remains to be seen what accounts for the necessity of *proportional* punishment.

First, it must be noted that any conception of the role of the justice system that is not based on discharging negative deserts will yield alternative justifications for punishment. In addition to the protection and security of others and fairness to victims, rehabilitation, deterrence, lowered recidivism, and threats to property offer themselves as possibilities. A criminal sentence can be designed to minimize harms or to maximize benefits. However, it would be extremely unsatisfying to qualify our HR with some kind of instrumental calculus. Instead, consider that all of the considerations above can also be formulated as HR held

by victims, society, or the offender themselves. All people have the HR to safety and security of person, and economic and social security and stability as well. When these rights are threatened by criminal activity, then the criminal's HR are in conflict with the HR of victims or potential victims. Effective rehabilitation might support an offender's right to health or self-determination; interestingly, in this case their right to freedom of movement can come into conflict with their own rights.

Once a conflict between HR has been established, an instrumental calculus mediating that conflict might be more appropriate. Alternatively, rights can be weighed against one another according to their relative normative force. For example, the right to life and the right not to be tortured are considered to be particularly forceful; they might outweigh most other rights. Couching the justification of punishment in terms of competing rights, mediated by instrumental concerns or relative normative force, gives rise to proportionality. In the first case, it would be needlessly cruel to punish someone if no good comes of that punishment; therefore, the appropriate sentence should be proportional to its aims and their value (lowered recidivism, deterrence, rehabilitation, safety and security of persons and property, fairness, etc.) In the second case, the normative force of rights can be thought of as setting the boundary where the domain of one right ends and the other begins (I will say more about this in Section V.). Moreover, it can be argued that judgments based on considerations of desert may be harder to implement fairly, are more resistant to standardization, and are less accessible to empirical inquiry than considerations stemming from forward-looking concerns like lowered recidivism or the future security of persons and property. Therefore, to the degree that attempting to ascertain an offender's desert basis results in an inappropriate punishment, HR protecting against such abuses might in fact limit considerations of desert in sentencing.

The question of what an inappropriate punishment might

be points to another problem with desert. Nickel's assertion that "Human rights requiring that punishments be proportional to crimes committed seem to be at least partially desert-based" (Nickel 2015, p. 160) is misleading. In fact, despite the language in the ECFR, I would argue that there is no HR to face a punishment proportionate to one's crime. The true object of the right is protection from punishments that are greater than what some measure of proportionality would entail. If that measure were tracking desert, and desert was undergirding the criminal justice system, it would be unjust *to defendants* for judges to give lenient sentences. If it were mediating between rights in conflict, then there is a justification for sentences that "feel" lenient in a desert-oriented sense—they are attempting to minimize harm or set appropriate boundaries between rights.

The rights in conflict approach has other benefits. For one, because it can be understood as *human* rights in conflict, it can preserve the special status that HR may hold in relation to other rights. More importantly, it allows for equal limitations on rights that have positive implications for the universality of HR. On the face of things, the fact that certain protections are lost when weighed against others might appear to compromise, as opposed to rescue, HR's universality. However, the important point here is that every person holds their rights equally. If all HR can be understood to extend up to the point where a conflicting HR has a stronger claim (or where it runs up against a non-derogable right[7], then all people have every right to the same degree—the same limitations hold equally for all people. According to desert-based accounts, individuals with negative desert bases would have greater limitations imposed on their rights, and the size of those limitations would be continuously shifting according to those bases.

Section V: Objections

In this section, I will address two possible objections to the rights in conflict account of proportional punishment.

Section 5.1: The Inalienability of HR

One possible objection is that, if an account of HR in conflict is integrated into a theory of HR, the possible inalienability of HR is lost. To use the example outlined in Section IV.2, if a person is incarcerated to protect the safety and security of their community, they lose the HR to freedom of movement. Therefore, that right is not inalienable. I believe the same kind of response can be made here that I made when universality appeared to be under threat. The right to freedom of movement extends only so far—it has inherent limitations built into its structure. To see this, consider that the right to freedom of movement does not give me the right to walk onto the private property of others. This limitation is understood to be baked into the structure of the right precisely because the right to private property and the right to freedom of movement conflict. It would be unreasonable to say that when I am told I cannot walk into my neighbor's house I have lost my right to move freely. Where the line is drawn corresponds to the relative force of the rights, which in turn is related to the interests at stake. If the government restricts my access to public property and institutions or prevents me from crossing international borders, my interests are severely threatened. On the other hand, nothing much is at stake by regulating my access to private property.

However, there is a difference between these two examples. In the latter, the limit on the right is fixed—the boundary between my right to freedom of movement and my neighbor's private property rights is stationary. In the former, the limit can move to accommodate the capriciousness of the government. While it could be argued that any right which could come into conflict with freedom of movement sets a fixed limit on what can reason-

ably be claimed by virtue of this right, a complete account of all conflicting rights and their limits in relation to one another would be impossible to formulate, and possibly too rigid to use effectively. Therefore, there is a certain amount of ambiguity involved when lines between rights are drawn (I will have more to say about this in Section V.2). This is a problem for implementing the rights in conflict approach in general, and perhaps a problem for inalienability in particular. The answer to the question of whether a right can be said to be inalienable if its boundaries are theoretically fixable but practically uncertain will perhaps differ from person to person. Certainly, it raises the specter of doubt and insecurity for the rights holder. For myself, it makes little difference, because I do not believe it is possible to come up with a theory of HR where no HR ever conflict. Consider the right to shelter and the right to healthcare, for example. For lack of resources, some states will have trouble delivering on both of these. And, while some theorists claim that some positive rights are too burdensome to require duties by state actors, and so cannot qualify as HR at all, canonical HR like due process also require considerable economic resources. If it is the case that mediating between these conflicts damages the inalienability of HR, impoverished states would still have a duty to do so somehow. And ultimately, qualifying HR on the basis of desert alienates individuals from their rights by definition, anyway.

Section 5.2: Designating Weak Rights as HR as a Category Error

Another objection might arise from a disagreement about the relative difficulty of making judgments based on desert and of mediating between rights in conflict. It could be said that courts routinely make decisions based on what offenders deserve, or what rights holders deserve. They are trained to do this, and it is not as uncertain and ambiguous a judgment as I have made it out to be. I

want to address this reply to individuals who might hold this view, and also to those who are committed to the role of desert in HR, either because their commitment to the idea that the possession of individual rights is always contingent upon personal deserts or their commitment to retributive justice outweighs their concerns over losing HR's universality. Whatever one's commitments, I believe that tying proportionality to desert in a theory of HR is inappropriate, and stems from a category error. To show how, I will offer an argument put forward by Verdirame (2015). We can draw a distinction between rights that are limited in the sense that freedom of movement is when I am told I cannot enter my neighbor's home, and rights that are limitable. By limitable, what I have in mind are rights whose boundaries are fluid, moving relationally according to some basis such as desert. The holder of a right that is limitable will be insecure in their possession of the right. There will be no stable expectations regarding how far their protections extend since they are subject to change and thus unpredictable. This kind of right is necessarily weaker than limited rights and generates correspondingly weaker duties.[8]

This is inconsistent with human rights practice. Verdirame points out that the various preambles to the canonical HR documents make the claim that HR are the "foundation of justice and peace in the world," that violations justify "rebellion," and that "disregard and contempt for human rights have resulted in barbarous acts which have outraged the conscience of mankind (Verdirame 2015, p. 352).[9] This tells us something about the importance of HR and the normative strength we should ascribe to them. As Verdirame puts it, giving a role to desert in the implementation of HR would create

> an inconsistency in the practice judged on its own standards of political justification. Those standards tell us that these rights are so fundamental that respect for them is what separates us from barbarism and war. If that is true,

> how can we justify making the enjoyment of some of these rights uncertain, vague and weak? (ibid, p. 352)

Put simply, limitable rights do not have sufficient normative force to rate as HR. To claim that HR can be weak rights whose protections can shift according to states' judgments is a category error. It sends a message to rights holders about how seriously they can take those protections and undermines international legal HR practice. Therefore, whatever limitations a successful theory of HR sets on specific rights in order to mediate between rights in conflict, they have to lead to stable expectations in practice. If they do not, then the rights they limit cannot be considered HR at all.

I do not see how desert considerations can be used to establish limited, as opposed to limitable, HR. Individual's desert bases change over time. There is no methodology for establishing desert's magnitude and no units for measuring it. It is a highly subjective concept (meaning that an individual's desert basis will change depending on who is measuring it), and its ambiguity leaves rights holders with little recourse to appeal when they are judged undeserving. Further, as I showed above, relying on considerations of desert undermines HR's universality. I admit that mediating between rights in conflict is theoretically challenging and can lead to other disagreements and ambiguities. However, the results can be strong, limited rights that yield concrete expectations. It seems to me that all that is necessary to confirm this is an appeal to intuition; although I may not know the boundaries of my rights with precision, I have a strong enough sense that I know that it is wrong to wander onto private property or to incite wanton violence under the guise of my freedom of expression. Negotiating these limitations is simply to negotiate what the rights are. Fuzzy boundaries are preferrable to unstable ones.

Section VI: Practical Concerns with Considerations of Moral Desert in the Implementation of HR

So far, I have focused on theoretical concerns with accepting considerations of moral desert as qualifying or justifying HR. I would like to say a few words regarding the deep misgivings I have about how desert can be and is used in this manner and on the repercussions of that usage. Generally, the duties corresponding to our HR are thought to be held by states. One reason for this is that, historically, states have been the perpetrators of the most egregious HR violations. However, if desert can qualify or justify HR, then it will be states (as the primary enforcers of international human rights law) and their institutions who will determine individual's desert bases according to whatever criteria they see fit. From these bases, states will decide which individuals should be excluded from receiving the objects and protections of some HR, and the degree to which other individuals will have some HR limited. I would hate to be in a position in which my HR were being violated and I had to assert a claim on the violating state, only to be told that my HR, which I have by virtue of my humanity, had been abrogated because I did not deserve those rights. It does not seem rational to trust states to make judgments on the basis of desert.

When HR are enforced, it is often done through inter-state interference. However, there are no international criteria to determine desert bases or the resulting qualifications or limitations. Judgments of moral desert are not universal across cultures. In fact, there is considerable controversy surrounding whether HR are truly universal in a cross-cultural sense, or if they reflect and impose Western values. Because of this, and because some HR have only been implemented selectively (such as the Palestinians' right of return as laid out in article 13 of the UDHR),[10] some people believe that the HR agenda is imperialistic in nature. The

legacy of colonialism offers many good reasons to be suspicious of claims of universality in this particular sense. And yet, at the very least it seems safe to adopt a consensus approach to the cross-cultural universality of HR; those rights that the vast majority of nations uphold as HR should be considered HR (unless the consensus was achieved by coercion.) Of the world's 195 nations, 193 have ratified the UDHR, and 172 have ratified both the ICCPR and ICESCR.

There is a tension between the rights in these documents, which are supported through nearly unanimous global consensus, and some state-held positions regarding how to determine the desert bases of individuals and what should follow from those bases. One example given by Jack Donnelly is the withholding of protections against violence towards homosexuals by states, or violence perpetrated against homosexuals by states. He says that

> [E]veryone is entitled to security of the person. If the state refuses to protect some people against private violence, on the grounds that they are immoral, the state violates their basic human rights which are held no less by the immoral than the moral. And the idea that the state should be permitted to imprison or even execute people solely on the basis of private voluntary acts between consenting adults, however much that behavior or "lifestyle" offends community conceptions of morality, is inconsistent with any plausible conception of ... individual human rights. (Donnelly 2007)

I agree with his assessment. If considerations of moral desert are embraced as a justifying or qualifying feature of HR, the doors are open for further transgressions.

Section VII: Conclusion

In conclusion, the role of HR practice and discourse continues to

grow in contemporary international politics. Theories of HR can help to distinguish between those rights that qualify as HR and those that do not. A good theory of HR would need to meet the following two criteria. It would have to preserve those features that are most important to our conception of HR, such as their universality, their essential nature, and their strong normative force. And, it would need to justify (at least most of) the current catalog of canonical HR whose implementation have had such a positive impact on the international political order and the lives of countless individuals. On the face of things, Nickel's suggestion that considerations of moral desert can justify HR, such as the right to proportionate pay and the right not to be punished disproportionately, feels intuitive and might seem to satisfy the second criteria. However, it undermines every feature of the first. I suggest that understanding proportional pay as a secondary right derived from the basal right not to be exploited justifies the proportionality inherent in that right more effectively. The HR-in-conflict approach does similar work justifying the proportionality inherent in the right not to be punished disproportionately. Both of these approaches leave the door open to a wide range of philosophical foundations, preserve the features of HR that many are committed to, and prevent HR abuses that could result from the implementation of desert considerations in HR practice.

Notes

1. Nickel defines proportionality in reference to desert. He says, “P1 deserves T from P2 in virtue of having DB P1 is a person who deserves something, T is some treatment or state of affairs that is deserved, P2 is a person who is permitted or obligated to give or impose T, and DB is the desert basis. The desert relation puts these elements together and asserts that T in its size and nature is permissible or required in light of DB for P2 to give to or impose on P1. This is the proportionality aspect of desert. Both T and DB must admit of degrees so that more or less of T can be proportional to more of less of DB" (Nickel 2015, p. 155). However, proportionality obviously does not have to include any reference to desert. Letsas (2015) defines four types of proportionality, two of which can help me describe what I have in mind. The first is mathematical, where proportionality serves to describe a “fixed relation-

ship between two variable quantities..." (Letsas 2015, p. 318). The second is normative. In one sense it is similar to mathematical proportionality, in that it tracks a relationship between two or more variables. However, instead of describing that relationship, it tells us what it ought to be and reflects some "moral ideal" (ibid, p. 318). For example, the number of seats each state has in the House of Representatives is proportional to the number of citizens of each state and should reflect the way democracy ought to operate. This kind of proportionality is what I am referring to in this paper, and what normative concept proportionality should reflect when it appears in HR is at the center of my disagreement with Nickel.

2. It quickly became clear after the ICCPR and ECHR were passed that many groups did not enjoy their HR equally. Subsequent conventions were passed to draw special attention to this fact, and to give those groups additional protections that mirrored their unique interests. Examples of these conventions include the Convention on the Elimination of All Forms of Racial Discrimination (1969), the Convention on the Elimination of All Forms of Discrimination against Women (1979), the Convention on the Rights of the Child (1990), the International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (2003), and the Convention on the Rights of Persons with Disabilities (2008).

3. It is not my intention to denigrate the concept of human dignity or dismiss it as a possible foundation of HR. In fact, insofar as it is sometimes identified with the intrinsic moral worth of human beings, I think it is very promising. However, even Waldron (2015), who supports dignity as a foundation of HR, recognizes problems with attempts to define the term, and points out that it has been referred to as "subjective," "squishy," and "a mere slogan." He suggests that "our understanding of its meaning is a work in progress" (ibid, p. 121). David Luban (2015), a proponent of a pragmatic, political conception of human rights, believes that dignity's flexibility is one of its strengths, allowing it to undergird all contemporary legal HR. He attempts to solve the definitional problem by claiming that dignity should be defined through its usage in international legal human rights practice, as opposed to serving as a foundation of HR. I disagree with his solution, but I appreciate his point.

4. It may be the case that proportionality is not playing a role at all. I will not explore this possibility here.

5. Stemplowska does not address how the conditions underlying the formation of a contract can affect the resulting contractual rights (if one party is coerced, for example). These conditions are intimately connected to my argument for the derivation of the right to proportional pay from the right to freedom from exploitation.

6. This way of describing how one right is derived from another is similar to the way Judith Jarvis Thomson (1975) explains the derivation of privacy rights from property rights.

7. Some rights, such as the rights to life and freedom from torture, are sometimes referred to as non-derogable rights. People who hold this view believe that it is never permissible to infringe upon these rights.
8. Nickel himself is of the view that considerations of desert have weak normative force. He makes the case that such considerations can still generate duties. However, my contention is that these duties cannot correspond to the kind of rights worthy of being called HR.
9. Verdirame is referring to the preambles of the UDHR, the ICCPR, and the ECHR.
10. I owe Dr. Mohammed Abed a debt of gratitude for this example, as well as his invaluable insight, comments, and support while I wrote this paper.

Bibliography

Donnelly, Jack. (2007) "The Relative Universality of Human Rights," *Human Rights Quarterly* 29(2) pp. 281-306

Gould, Carol C. (2015) "A Social Ontology of Human Rights," in: Cruft, Rowan, Mathew Liao, & Massimo Renzo (Eds), *Philosophical Foundations of Human Rights*, pp. 177-195 (Oxford: Oxford University Press)

Griffin, James. (2008) *On Human Rights* (Oxford: Oxford University Press)

Letsas, George. (2015) "Rescuing Proportionality," in: Cruft, Rowan, Mathew Liao, & Massimo Renzo (Eds), Philosophical Foundations of Human Rights, pp. 316-340 (Oxford: Oxford University Press)

Liao, Mathew. (2015) "Human Rights as Fundamental conditions for a Good Life," in: Cruft, Rowan, Mathew Liao, & Massimo Renzo (Eds), *Philosophical Foundations of Human Rights*, pp. 101-116 (Oxford: Oxford University Press)

Luban, David. (2015) "Human Rights Pragmatism and Human Dignity," in: Cruft, Rowan, Mathew Liao, & Massimo Renzo (Eds), *Philosophical Foundations of Human Rights*, pp. 263-278 (Oxford: Oxford University Press)

Miller, David. (2012) "Grounding Human Rights," *Critical Review of International Social and Political Philosophy* 15(4), pp. 407-427

Nickel, James W. (2015) "Personal Deserts and Human Rights," in: Cruft, Rowan, Mathew Liao, & Massimo Renzo (Eds), *Philosophical Foundations of Human Rights*, pp. 153-165 (Oxford: Oxford University Press)

Nussbaum, Martha. (2011) *Creating Capabilities: The Human Development Approach* (Cambridge, MA: Harvard University Press)

Shannon, Sarah K. S., Christopher Uggen, Jason Schnittker, Melissa Thompson, Sara Wakefield, and Michael Massoglia. (2017) "The

Growth, Scope, and Spatial Distribution of People with Felony Records in the United States," *Demography* 54(5), pp. 1795-1818

Tasioulas, John. (2015) "On the Foundations of Human Rights," in: Cruft, Rowan, Mathew Liao, & Massimo Renzo (Eds), *Philosophical Foundations of Human Rights*, pp. 45-70 (Oxford: Oxford University Press)

Thomson, Judith Jarvis. (1975) "The Right to Privacy," *Philosophy & Public Affairs* 4(4), pp. 295-314

Verdirame, Guglielmo. (2015) "Rescuing Human Rights from Proportionality," in: Cruft, Rowan, Mathew Liao, & Massimo Renzo (Eds), *Philosophical Foundations of Human Rights*, pp. 341-360 (Oxford: Oxford University Press)

Waldron, Jeremy. (2015) "Is Dignity the Foundation of Human Rights?" in: Cruft, Rowan, Mathew Liao, & Massimo Renzo (Eds), *Philosophical Foundations of Human Rights*, pp. 117-137 (Oxford: Oxford University Press)

A Question of Control

Heather Norwood

Introduction

Do humans have free will? Do we have the freedom to really do *anything*? Is the world open to us, like a buffet of choices, from which we can load our plates with whichever ones we want? Or are our choices determined, and everything we do controlled by fate? According to Hume, our choices are determined, but they are determined by our desires, and in that way, we are free. There is little doubt that we have this Humean compatibilist free will—everyone generally experiences making choices in accordance with their desires—but incompatibilist views argue that this kind of free will is not free will at all because free will is not compatible with our choices being completely determined. In this paper, I will focus on just the libertarian answer to the question of free will, specifically as discussed by Mark Balaguer in his book *Free Will as an Open Scientific Problem* (2010), and even more specifically, just Chapter 3: Why the Libertarian Question Reduces to the Issue of Indeterminacy.

Libertarianism, as Balaguer defines it, is the view "that human beings possess L-freedom, where a person is *L-free* if and only if she makes at least some decisions that are such that (a) they are both undetermined and appropriately nonrandom, and (b) the indeterminacy is relevant to the appropriate nonrandomness in the sense that it *generates* the nonrandomness" (Balaguer 2010, p. 65). On the libertarian view, a person has L-freedom if there are choices that she makes that are not predetermined by her desires, by fate, or by some causal process that results in her having no choice but to do what she does, and this lack of determination produces or gives rise to a kind of nonrandomness

such that she has both authorship and control of the outcome of her choice. These kinds of appropriately nonrandom choices are further characterized by a lack of coercion, force, or any kind of external control that would make one question whether or not the choice was really *hers*.

As the title of the chapter suggests, Balaguer reduces the question about whether humans have L-freedom to a question about whether a certain kind of indeterminism is true: TDW-indeterminism. TDW-indeterminism is concerned with a kind of choice, namely "torn decisions," which Balaguer defines in terms of their phenomenology: a torn decision is one "in which the person in question (a) has reasons for two or more options and feels torn as to which sets of reason is strongest, that is, has no conscious belief as to which option is best, given her reasons; and (b) decides without resolving this conflict—that is, the person has the experience of 'just choosing'" (ibid, p. 71). TDW-indeterminism is a claim about torn decisions and is defined as follows:

> Some of our torn decisions are wholly undetermined at the moment of choice, where to say that a torn decision is wholly undetermined at the moment of choice is to say that the moment-of-choice probabilities of the various reasons-based tied-for-best options being chosen match the reasons-based probabilities, so that these moment-of-choice probabilities are all roughly even, given the complete state of the world and all the laws of nature, and the choice occurs without any further causal input, that is, without anything else being significantly causally relevant to which option is chosen (ibid, p. 78).

Balaguer's claim is that *if* torn decisions are undetermined (in a TDW-indeterminism kind of way) at the moment of choice *then* (a) they are not just undetermined but appropriately nonrandom and (b) the indeterminacy procures the appropriate nonrandomness, and *if* (a) and (b) are both true, *then* libertarianism is also

true (ibid, p. 68).

In Section I, I present an objection to Balaguer's claim, namely Peter van Inwagen's rollback argument, and Balaguer's responses to the issues of chance, luck, and randomness that the argument raises. In Section II, I evaluate the first two points that Balaguer gives as an answer, defending the libertarian view as much as possible, while grappling with the issues of chance and luck and considering the role of causation. Section III then focuses simply on randomness, and illustrates how Balaguer's third and fourth responses to chance or luck objections do not lead to either (a) or (b). Finally, in Section IV, I offer additional senses of randomness that procure more control for the agent with or without indeterminacy, thus illustrating that Balaguer's if/then claim is false—that the indeterminacy in question does not procure or generate the appropriate nonrandomness because the indeterminacy leads to a kind of inescapable randomness which diminishes control.

Section I: The Rollback Argument

Peter van Inwagen (2000) argues that indeterminism and free will are incompatible, and claims that indeterminism, rather than being freedom-enhancing, or procuring appropriate nonrandomness, instead is freedom undermining and procures chance. Van Inwagen imagines a scenario where Alice, who, when faced with a torn decision of whether to lie or tell the truth, chooses to tell the truth. God subsequently rolls back Alice's choice a thousand times, so that we get a thousand replays of Alice's decision. In this series of replays, based on roughly even probabilities associated with Alice's options (and as given to us in the definition of TDW-indeterminism), we should see her choose to lie roughly five hundred times and choose to tell the truth roughly five hundred times. The problem with this, according to van Inwagen, is that Alice's choice resolves into a matter of chance—if on the seven-

hundred-and-twenty-seventh replay "each of the two possible outcomes of this replay has an objective 'ground floor' probability of 0.5… then in the strictest sense imaginable, the outcome of the replay will be a matter of chance" (van Inwagen 2000, p. 15). The implication of van Inwagen's scenario is that if a choice is undetermined (like a torn decision) and has these roughly even probabilities, then it is *chance*, and not the agent's free will, that determines which choice is made. And if this is the case, then Alice does not author and control her decision to tell the truth in a way that is appropriately nonrandom, so that the indeterminacy in question does not procure non-randomness.

Balaguer addresses this objection directly, making four points, which I repeat here, in relation to Alice.

(1) There is no reason to believe that there is any inconsistency between the following two claims: (i) Alice can choose differently in different replays of the decision, and (ii) in each of these replays, it is Alice who does the choosing, or who authors and controls the decision. This means that even though Alice may choose to lie on one replay of the decision, and tell the truth on another replay, any event of her either telling the truth or telling a lie, is an Alice-choosing event;

(2) Since lying and telling the truth are equally weighted in Alice's head, we should expect that, given different plays of the decision, she will choose different things. That is to say, it is not a problem that Alice chooses differently in different replays, because if she did not, then we would have reason to think that there was a hidden cause, one that always resulted in Alice making the same choice every time (and in that case the outcome would be determined);

(3) That the choice was made randomly or arbitrarily (or as a matter of chance or luck) in *some* sense does not undermine the fact that Alice made the choice and therefore she was the one who authored and controlled which option was chosen—

and, hence, that it was non-random in the sense that matters to free will; and

(4) The fixed, roughly even, probabilities related to a torn decision do not result in an outcome that is a matter of luck because the decision in question is still made *by Alice.*

The upshot of these four responses is this: "just because [the] decision is arbitrary or random (or, if you like, chancy or lucky) in *some* senses of these terms, it doesn't follow that it was arbitrary or random in the sense that's relevant here" (Balaguer 2010, p. 94) (i.e. in the sense of nonrandomness which is required for free will, specifically in the senses of authorship and control). Balaguer is arguing, contrary to van Inwagen, that the indeterminacy in question may seem to be chancy, but that it does not undermine the fact that the agent is the one who makes the choice and who controlled which option was chosen. In other words, Balaguer would say that even though Alice's choice to tell the truth may have some chance aspects to it, that it was nonetheless an Alice-choosing event: the choice flowed from her in a way that allowed her to maintain authorship and control over what choice was made, and in that way the choice was appropriately nonrandom.

Section II: Chance, Luck, and Predictability

In this section, I will evaluate counter-points (1)–(3) to the rollback argument that Balaguer offers, as it *does* seem to be the case that because we are unable to predict Alice's choice, that the outcome of her choice is a matter of chance. It also seems to be the case that if we could say, conversely, that Alice caused the outcome, then we could overcome chance and Alice would have control over her choice in a torn decision scenario. Balaguer's libertarian view, in order to be plausible, needs to offer a concrete explanation of how Alice has control in the midst of random outcomes.

The first point that Balaguer makes is that there is no reason

to doubt that (i) Alice can choose differently in different replays of the decision and (ii) in each of these replays, it is Alice who does the choosing, or who authors and controls the decision. Part of this response seems undeniably true: it is possible for Alice to choose differently on different replays, and that each time she chooses, regardless of the choice she makes, it is her that is making the choice. This fact, that the choice always belongs to Alice, means that she definitively authors each choice, regardless of the outcome. However, the issue that van Inwagen brings to light is not one of authorship, but one of *control*. It is not clear, in every instance of Alice choosing, that she is in control of which outcome is chosen just because she is the author of that choice (which means nothing more at this point than that it is hers). For example, she is not the author of the choice in the way that one is an author of a book, in which each word was consciously placed in its spot. Rather, because this is a torn decision scenario, Alice makes a conscious, but arbitrary or random, choice. She does not make the kind of choice that is a result of a conscious deliberation where one tied-for-best option is chosen—if this were the case, by definition, this would no longer be a torn decision scenario for Alice. Due to this fact, Alice authors the choice, but instead of her consciously placing each word in its place, in this section of the book, Alice closes her eyes and taps on the keyboard randomly and "just types," saving whatever comes out. Are we then to say that Alice consciously controlled what she "just typed"? If we accept Balaguer's first point, it seems that we would have to say yes—that Alice, when she closed her eyes and randomly hit the keys, was both authoring and controlling what ended up in that section of the book. However, this seems like an odd thing to say because it seems like whatever she typed out when she was "just typing" was a matter of chance.

However, Alice *did intend to make a choice*, and I think that Balaguer's first point pulls on this intuition. Perhaps Alice did not consciously choose to lie or tell the truth, but she did choose to

choose (or to close her eyes and hit some keys), and there is some sense in which this does account for the chanciness related to the outcome. It is not by luck or chance that the choice itself occurs, and it is not by luck or chance that the outcome for Alice is either choosing to lie or to tell the truth, because the choice is Alice's and she intends to make it, and all other options have already been ruled out by Alice's conscious reasons.

The second reason that it is not a matter of chance or luck is because of the way that the probabilities are involved: assuming, as Balaguer does, that the probabilities are roughly even, the outcome is not probabilistically determined in a chancy kind of way. There is a connection between predictability and chance, as we have already noted, and predictability is related to probabilities, because we specifically use probabilities as a way to predict outcomes and this is exactly why we say that an outcome was lucky if the chances of that outcome occurring were low. If the chance of the outcome is low that means that the probability of the outcome occurring was low, but in the case of Alice, the chances of either outcome occurring are roughly even, so in this sense, the outcome is not a matter of mere chance, and it is not lucky that Alice chose to tell the truth. The outcome is specifically probabilistically undetermined. However, what remains is that the outcome was random, because nothing determined which outcome occurred—it wasn't Alice, because Alice "just chose" without coming to a conclusion about which option was best, and it wasn't chance, because the probabilities were roughly even. Since the outcome was undetermined, the choice between her tied-for-best options was arbitrary. The fact remains that the outcome for Alice is random because *her choice is random*.

Balaguer's third point is that we should expect Alice to choose differently in different replays of her torn decision, and that if she does not, there must be some underlying cause for the same outcome occurring each time the choice was played out. Analogously, if we roll a pair of dice a thousand times and they

come up snake eyes every time, we will think the dice have been weighted in some way (i.e., that there is some kind of cause) that determines the outcome. This pulls on something that we want out of the concept of causation—we want to say that when something is caused, we can predict the outcome, and if it is not caused, we cannot predict the outcome. Conversely, if an outcome can be predicted, it is easy to say that it is caused (and generally we are able to say as a part of the prediction, what is causally involved in an outcome), and if it cannot be predicted, then it is not caused (or causally determined). With this in mind, van Inwagen's argument can be read as working backwards from prediction: if you cannot predict the outcome of Alice's choice, then the outcome was not determined by her, and if it was not determined by her, then it was not caused by her, and if it was not caused by her, then she was not in control of the outcome, and therefore, it was a matter of chance, and not free will that she chose to tell the truth rather than to lie. Van Inwagen here is playing the part of Anscombe's hypothetical physicist who "hates a theory… that essentially assigns only probability to a result, essentially allows a range of possible results, never narrowed down to one until the event itself" (Anscombe 1981, p. 142).

It seems that underneath the rollback argument, there is an issue of causality: without causation, there's no control. Further, it seems that torn decision scenarios embody a theory akin the one described above—in a torn decision (as defined), there are only probabilities and a range of outcomes and no one, not even the person making the choice, knows the outcome of the event until it occurs. The kind of causality that Anscombe offers, indeterministic causality, precludes one from being able to predict an outcome based on a state of affairs or an event together with the laws of physics, and points to a lack of control on the part of the agent, even in the midst of causality. Why would we want a causality that is deterministic? Because deterministic causality answers the question of control—if we could say that Alice had

control over which outcome occurred in her torn decision scenario because she *caused* it, we could easily do away with chance and luck objections. However, if the only kind of causality present is indeterministic causality, then the question of control remains.

As Griffith (2010) points out, there "are strong intuitive connections between luck and chance," but these connections are interwoven with predictability and arbitrariness, so we can say, without sounding illogical, that "if something is theoretically unpredictable, there is a sense of arbitrariness such that its occurrence feels like a matter of luck" (Griffith 2010, p. 44). I think this is correct, and it aligns with an intuitive sense of causation as deterministic with predictable outcomes that, when present, cancels out the factors of chance and luck. That is to say, when we confirm that something is caused or determined or causally determined, we can definitively conclude that it was not a matter of chance or luck. It seems that van Inwagen's argument (like all arguments of the sort) demands something that cannot be found: deterministic causation that results in a predictability that satisfies an apparent condition for control. But we should not expect to find this on the libertarian view because, by definition, it requires something besides deterministic causation (i.e., indeterministic causation or non-causation) in order to be true.

So, if we take the libertarian stance, we can respond to van Inwagen by saying that part of what he requires for free will *is* present in the libertarian view: causation. There is causation in the Alice scenario, but it is of the indeterministic variety, and we cannot expect to be able to predict the outcome of Alice's choice because of the undetermined nature of events that still holds on a view of indeterministic causation. This is a good libertarian response, and blocks part of the argument by providing for causation in the absence of predictability. If we understand indeterministic causation to be part of the reasoning behind Balaguer's second point, then it is a good rebuttal to this aspect of the luck objection in the rollback argument, as it actually does provide an answer for

part of the issue. We should, as (2) says, see Alice choose differently on different plays because the event is undetermined, and the causal process in motion is indeterministic, which just means that on different replays, there will be different outcomes.

Section III: The Inescapable Randomness

We have now answered issues of luck, chance, causation, and predictability. What remains is randomness, and Balaguer's third point takes this on directly, asserting that even though the choice was made randomly or arbitrarily in *some* sense, it does not undermine the fact that Alice made the choice and, therefore, she was the one who authored and controlled which option was chosen. Balaguer does admit that there are multiple senses of randomness at play in torn decisions, and claims that the sense of randomness that we should care about is the sense of randomness that matters in the case of authorship and control. He answers the randomness charge with the agent herself—with Alice. The choice is not random because Alice makes the choice, and because she makes it, she authors and controls the choice. However, that Alice makes the choice does not cancel out another sense of randomness here: that the outcome is random. Because of this, it appears that Balaguer's response only supplies enough non-randomness through Alice that we can say that she is the author of her decision, but due to the decision itself remaining random, it is still difficult to see how Alice is in control of her choice. This is what I see as a kind of inescapable randomness; even when we grant Balaguer's points (1) and (2) against van Inwagen, we are left with an arbitrary choice and a random outcome, and point (3) does not really offer a satisfactory answer to the issue of control.

I would argue that, if torn decisions are undetermined in a TDW-indeterminism kind of way, then we do not get enough appropriate non-randomness to say that the person making the decision is in control of which outcome occurs. What seems to

have happened is that, when we shift away from issues of chance and luck, predictability and probability, we are left with randomness because of the continual presence of indeterminacy. What remains is the question of control. It is counterintuitive to think of a random outcome as controlled. If there is a range of possible results, and the outcome isn't determined until it happens, then the outcome is still random. But we are not really concerned with random outcomes, we are concerned with random decisions—with this idea that *Alice chose randomly*—because the issue is whether Alice is in control of her decision. The libertarian claims Alice is able to choose which outcome occurs through her own free will. But indeterministic causality, which is the main property of the Alice-choosing event, does not appear to allow for this possibility at all. Alice "just chooses" without having come to a conclusion about which tied-for-best option is really the best option. Thus isolated, we see something important about randomness: the concept of libertarian freedom turns on non-randomness because randomness is the enemy of control, and it remains even after we set aside chance and luck.

Recalling Balaguer's four responses to the rollback argument, we can see that authorship is never the issue; at every turn Alice is making the decision and Alice is the author of that decision because it is hers. However, that it is hers does not seem to lead definitively to the fact that she controls which outcome is chosen. It seems to be part of both the definitions of torn decisions and TDW-indeterminism that she cannot determine the "what" of the decision in a conscious way—the decision between tied-for-best options will always be arbitrary, and will always carry a sense of randomness because of that fact. It does not seem that there is any way around that sense of randomness—it is inescapable. Simultaneously, it also seems that it is this a sense of randomness which precludes control.

Section IV: The Randomness that Isn't

So, we have a sense of randomness that is a lack of control, as well as the senses of randomness that are related to unpredictability and causality (which we have attempted to resolve), but there are different senses of randomness besides these. There is a sense of randomness related to contrastive reasons (see Griffith) where we say that something is random if we cannot provide some kind of explanation for its occurrence. There is also a sense of randomness to an event that just happens to someone, like winning the lottery. But there are other senses of randomness still.

Consider this: I was vacuuming in the hallway, and behind me was a little table covered in sea shells, with a box of small trash bags on the corner, which I knocked off the table with the vacuum handle. The box fell onto the canister of the vacuum, hitting the power button hard enough to turn off the vacuum. When the vacuum turned off, I expected to see the cord of the vacuum pulled taut, because the cord is short and I often vacuum until I have pulled the cord taut enough to dislodge the plug from the outlet and shut off the vacuum. However, the cord was still loose, and upon seeing the box on the floor I realized that it was the box falling onto the power button that shut off the vacuum. This scenario is chancy in that the outcome was not predictable by probabilities, and lucky, since the probabilities of the outcome were actually very low, and it was also random in the sense that I had no idea it was coming. In fact, I was surprised to realize the trash bag box had turned off the vacuum after I knocked it off the table unknowingly. The sense of randomness that is highlighted here is the sense of surprise that comes from an event occurring which is completely unexpected. Additionally, there are senses of randomness inherent to this event which are related to a lack of a pattern or order. It's not the case that every time I vacuum I knock something onto the vacuum that shuts it off. In fact, I usually unplug the vacuum by pulling it too far, so if that had been the

reason the vacuum had shut off, it would have been a part of the pattern of what always happens when I vacuum and in that way, it would have been non-random. It is also not the case that the box falling event that resulted in the vacuum shutting off event was a part of a series of events that constitutes an instance of vacuuming. Based on what occurred both before and after the vacuum being shut off by the box, we have no reason to conclude that that event was one that was necessitated in some way, not even if we say that my knocking the box off the table was a necessitating event, because that only necessitated that the box would fall, not that it would fall onto the vacuum, definitely not that it would fall on the power button, and certainly not that it would fall on the power button in such a way that it would shut off the vacuum. The lack of necessitation points to a lack of order involved in the occurrence of the event of the vacuum shutting off and in that way the event is random in yet another sense.

So now we can characterize random events as: unpredictable, uncaused, uncontrolled, without reason (unexplainable), something that happens to someone, unexpected, without order, and without a pattern. If we consider Alice, we can ask: how many of these senses of randomness truly apply to her undetermined choice in a torn decision scenario?

We should not expect Alice's decision to be predictable because of the undetermined nature of the Alice-choosing event, but there is a sense in which her decision is partially predictable by Alice herself. If you asked Alice in the midst of her torn decision what she was about to do, she would tell you that she is either going to tell the truth or tell a lie; Alice will tell you that she is going to choose between her two options, two options which she consciously has and of which she is consciously aware. In this way, Alice herself can partially predict the outcome of her own decision—it will be one of two options that have already been picked out by her reasons (both conscious and subconscious). Alice herself will not be surprised by her choice to tell the truth—

she knew that this option was a possible outcome—unlike the way in which I was surprised by the box shutting off the vacuum. Additionally, she would be able to give you reasons for either option, explaining why they are equally good (or equally bad) in her estimation. This shows that her decision is not without order, or without a pattern, because the events leading up to her torn decision necessitate that she will choose one of two options. Further, she will choose in a way that is familiar to her because it is plausibly always the case that when Alice faces a torn decision, she has reasons that she can explain and picks between two equally weighted options. This makes it arguably the case that the Alice-choosing event is not something that happens *to* her, but something in which she is an active participant. So, the answer appears to be that none of the above senses of randomness apply to Alice's choice in any given replay.

Conclusion

We can see that there is much more non-randomness related to Alice's decision than there is randomness. If we think of a continuum which has on the one side control and non-randomness, and on the other side a lack of control and randomness, then the less non-randomness there is related to Alice's choice, the more control she has over it. But does the indeterminacy in question procure these senses of non-randomness? It doesn't seem so. It seems to be the case that these senses of non-randomness are procured by the fact that it is Alice who is making the choice, which if this is the case, falsifies Balaguer's initial if/then claim. The indeterminacy leads to inescapable randomness which precludes control, and Alice procures for herself as much control as she could possibly have through other senses of non-randomness, senses which upon reflection, are present in all of Alice's choices, because they are hers.

Bibliography

Anscombe, G.E.M. (1981) "Causality and Determination," in: Anscombe, G.E.M., *Metaphysics and the Philosophy of Mind*, pp. 133-147 (Minneapolis, MN: University of Minnesota Press)

Balaguer, Mark. (2010) *Free Will as an Open Scientific Problem*. (Cambridge, MA: MIT Press)

Griffith, Meghan. (2010) "Why Agent Caused Actions Are Not Lucky," *American Philosophical Quarterly* 47(1), pp. 43–56

Van Inwagen, Peter. (2000) "Free Will Remains a Mystery: The Eighth Philosophical Perspectives Lecture," *Philosophical Perspectives* 14, pp. 1–19

Hutto and Myin Fail The Hard Problem of Content

David Dixon

Introduction

Most scientists and philosophers are committed to the claim that many of our mental states are *about* the world. In other words, they have *mental content*. This fact has so far stubbornly resisted satisfactory naturalistic explanation. One seemingly promising approach explains mental content by appealing to physical and biological information. For such an approach to succeed, however, it must overcome what Hutto and Myin (2013) call the hard problem of content (HPC). This problem arises from attempting to explain mental content in terms of information defined as the covariance of states (i.e., covariance information). Unlike other phenomena that can be reduced to the properties of their constituents, the properties of mental content are not reducible to covariance information. Hutto and Myin (2017) claim to have a solution: explain the natural origins of content (NOC) by contextualizing information as covariance in the evolution of human communication in early sociocultural interactions.

Does Hutto and Myin's solution work? In this paper, I conclusively show that it does not. Rather than provide a clear account of NOC that passes HPC, Hutto and Myin outsource the job to others, then fail to put those sources together in a satisfactory way. On their own, the principal philosophers they put to work—Bar-On and Priselac (2011) and Sterelny (2010)—come up short (and do not claim otherwise). On top of this, a major obstacle to charitably assessing Hutto and Myin's view is that their big claims are presented with a frustrating lack of organiza-

tion. Therefore, the main accomplishment of this paper is to bring some organization to that view, if only to dismiss it.

In Section I, I lay the required conceptual framework needed to grasp the problems of and challenges to NOC, as Hutto and Myin conceive them. Section I consists of three sub-sections devoted to developing Hutto and Myin's understanding of mental content, information, and communication. Then, in Section II, I present HPC. This sets up a test that Hutto and Myin's approach must pass if it is to be successful. I begin Section III by prescribing three additional conditions that Hutto and Myin's view must satisfy if it is to successfully explain NOC and pass HPC. After this, Section III breaks down into three sub-sections. I first set about synthesizing Hutto and Myin's view into a coherent presentation. This view is composed of two parts: Ur-intentionality and the concept of sociocultural scaffolding. In Section 3.1, I develop Hutto and Myin's view of Ur-intentionality in detail and point to some serious problems. Since Hutto and Myin appeal to Bar-On and Priselac for an augmentation of Ur-intentionality that is supposed to take us closer to mental content, in Section 3.2 I analyze their concept of expressive behavior. Lastly, in section 3.3, I sketch Sterelny's notions of human niche construction and "scaffolding." Hutto and Myin claim that sociocultural scaffolding can support the natural emergence of mental content. In this final section, I conclusively show that, on their view, it does not. Since Hutto and Myin's objective is a naturalistic explanation of mental content, I now begin by elucidating that notion.

Section I: The Terms of NOC

Section 1.1: Intentionality and Mental Content

Intentionality, the feature that mental states like perception and thought, are *directed at, refer to, stand for*, or are *about* things besides themselves, is an essential feature of minds. For comparison, consider a rock and a visual perception. The rock has phys-

ical features such as mass, density, gravity, velocity, and energy. Presumably, from a naturalistic view, visual perceptions also have physical features, such as a relationship with the activity of photoreceptors in the eye. But unlike rocks, mental states can have features that seem to include the features of other things. So, for example, in visually perceiving a cat, the mental state of seeing is, in some sense, directed at an object (e.g., a cat) that is not part of the mental state.

The intentionality of mental states isn't restricted only to existing concrete objects like cats and rocks. Our thoughts often center on things which are removed from us either in space or time, and often on things which do not exist. For instance, I may think of my cat though it is not in the room; I may think of Descartes though he is no longer alive; and I can think of Sherlock Holmes even though he never existed.

Though there are many ways in which intentionality manifests itself in our mental lives, to illustrate the notion of mental content that Hutto and Myin target, I will focus only on the notion of *singular thought*. Following Crane (2011), I understand a singular thought to be "a thought which is directed at just one object" (Crane 2011, p. 21). The thought that "my cat is cute" is an instance of singular thought. In Crane's hands, the term is used for both an instance of thinking and for "*what* is thought" (ibid, p. 22). It is the latter sense that Hutto and Myin recognize as the mental content of a thought. For them, a mental state has content if and only if it has the following features:

(a) *It refers or purports to refer.* Reference is generally thought to be a semantic notion with two senses. The first is a relationship whose relata include a token representation and a referent. The relationship between a name (e.g., "Sméagol") and its bearer (e.g., my cat) is an instance of reference in language. Hutto and Myin assume that thoughts also represent their referents. The second sense is the act of using a

representation to talk or think about a particular thing. When I think that "Sméagol is hungry" I am using the linguistic sign "Sméagol" to think about my cat;

(b) *Has propositional form*. Propositional form mirrors the syntax of a simple English sentence. The latter is composed of a subject term (e.g., "Sméagol") and a predicate term (e.g., "is hungry"). The subject term may be a proper name, definite description, pronoun, or demonstrative (e.g., "this" or "that"). The predicate term may be a noun or noun phrase, or a verb or verb phrase;

(c) *Has truth conditions*. Thoughts with propositional form represent their referents as being a certain way (i.e., they make a claim about them). In the thought "Sméagol is hungry," Sméagol is being represented as being in a state of hunger. As such, this thought may be true or false. It is true if and only if he is hungry, otherwise it's false.

The challenge of NOC is to give a scientifically oriented explanation of mental content in non-intentional terms. One approach to doing so is to explain mental content (and intentionality, in general) in terms of physical and biological information. I turn now to define the relevant notions of covariance information and semantic information.

Section 1.2: Information

Picture a perfectly white wall without any blemishes. The surface is completely flat. Looking at the wall, you can't distinguish any part of it as different from any other part. Now imagine a black smudge on it. This break in the uniformity of the wall is a *datum*. A datum is defined by Floridi (2010) as a distinction between two variables x and y, where the relationship of 'being distinct from' is left open to interpretation. In the example above, x is the white wall and y is the smudge, and x is distinct from y by the color,

shape, and size of *y*. In seeking NOC, the relevant data is a lack of uniformity in an environment to which an organism is or can become biologically sensitive. For example, *E. coli* is biologically sensitive to the sugar ribose in virtue of a transmembrane receptor protein called Trg.[1]

In everyday settings, the concept of *information* is usually understood as *semantic information* (**S**-information). For example, when we say that a newspaper provides us information about national affairs, we are referring to **S**-information. Following Floridi (2010), we will say that ***v*** is an instance of **S**-information if and only if (i) ***v*** is composed of ***n*** data (where $\boldsymbol{n} \geq 1$); (ii) the data conform to the rules or syntax that define well-formedness in their representational medium; and (iii) ***v*** is meaningful. Though condition (iii) refers to semantic content (i.e., what a sentence conveys), for our purposes, I will regard semantic content as parasitic on mental content. Thus, for mental states, I identify the meaningfulness of **S**-information with mental content.

By contrast, consider the common litmus test used in cleaning pools. Litmus is a mixture of dyes extracted from lichens. When litmus is exposed to acidic conditions, it turns red; when it is exposed to basic conditions, it turns blue. Pool cleaners use litmus because these color changes covary with the pH level in water. While this covariance of states provide information, the chemical states and reactions, themselves, do not have the features of conditions (ii) and (iii). The exchange of hydrogen ions does not conform to a syntax, as the English language does. Nor is that exchange itself something that possesses reference, propositional form, or truth conditions. The information provided by the covariance of states is not **S**-information.

In nature, covariance information (**C**-information) is ubiquitous. For example, the number of tree rings covary with a tree's age in years; the orientation of a daisy covaries with the position of the sun in the sky; the motility of *E. coli* covaries with concentrations of ribose. To appropriately sensitive interpreters, such examples

carry information in the following sense. **C**-information is defined as two states or systems, *x* and *y*, that are coupled together in such a way that *x*'s being **F** is correlated with *y*'s being **G**. While the data of **C**-information have no syntax, they are related to each other by physical constraints and processes. Thus, in the case of litmus, color change is governed by the laws of chemistry. In biology, the relata of a covariance relationship may reflect other constraints, too. The most salient constraint to Hutto and Myin is the process of evolution through natural selection. Specifically, the evolution of behavioral responses to features of an organism's environment or other's signals. As we'll see in the next section, the latter is also governed by the costs associated with reliable communication.

With these notions introduced, we can further refine the challenge of NOC in informational terms. Namely, it is the task of the information-based approach to explain the emergence of **S**-information in terms of **C**-information. Since Hutto and Myin focus on the relationships between organism's and their environments, including other organisms, I now proceed to outline the important notions of environmental cue and biological signal.

Section 1.3: Cues and Communication

Have you ever wondered how mosquitoes find you in the night? It turns out that mosquitos possess carbon dioxide receptors on their antennae that help them find you. Omer and Gillies (1971) found that when CO_2 was added to a wind tunnel in short bursts that resembled the pattern of inhalation and exhalation in mammals, mosquitoes moved rapidly up the tunnel toward the source.[2] Further, it was found that when mosquitoes encounter CO_2 in natural settings, they respond by flying upwind. This interaction, between the mosquito and particular concentrations of CO_2 in the air, is an example of a *cue*.

Following Maynard-Smith (2003), I define a cue as "any feature of the world, animate or inanimate, that can be used by an

animal as a guide to future action."[3] Importantly, in our context, the sense of the word "guide" does not imply any inferential thinking or mental content. Moreover, animals do not know, for instance, that *that cue means I should do behavior* **b**, because such knowledge requires mental content. Rather, for our purposes, a cue is understood as any environmental feature that correlates with a behavior that has evolved because it confers a net evolutionary benefit to an organism. In the case of our mosquito hunting for a food source, the appropriate presence of CO_2 (**F**) in the air (*x*) covaries with the fact that a mammal (*y*) is nearby (**G**). **F***x* is a cue for the mosquito insofar as its sensitivity and response behavior have evolved because of the net evolutionary benefit that the **C**-information—**F***x* carries the information **G***y*—confers. It is in this way that organisms "exploit" **C**-information from cues in their environment.

Before moving on, it is worth noting that Hutto and Myin have a teleofunctional view of the evolution of cues and signals. *Teleofunctionalism* is a view developed largely in the context of teleosemantic theories of mental content.[4] It posits that certain biological traits (e.g., sensitivity to CO_2) have a "proper function." These are viewed in teleological terms, i.e., in terms of what they are *for*. Thus, in mosquitoes, sensitivity to CO_2 has the proper function of being caused by certain concentrations of CO_2 and causing certain response behaviors. It is for this reason, according to Teleofunctionalism, that such traits evolved through natural selection. Lastly, Teleofunctionalism also has a normative aspect: there is something that a trait *should* do (i.e., its proper function) though it may fail.[5]

Returning, now, to the issue of how organisms use **C**-information to their benefit, Hutto and Myin are mainly concerned with the coevolution of animal signals in communication. This becomes clear once we consider NOC as the problem of how mental content might have evolved in the context of human communication needs. The way in which Hutto and Myin under-

stand such communication comes from Godfrey-Smith (2014) and Maynard-Smith (2003). Both frame animal communication in terms of senders transmitting signals to receivers. And both recognize the "strategic cost" of a signal as a selection mechanism through which signals can coevolve.[6]

A *signal*, according to Maynard-Smith, is "any act or structure which alters the behavior of other organisms, which evolved because of that effect, and is effective because the receiver's response has also evolved" (2010, p. 3). If the signal alters the behavior of its receiver, the signal must be beneficial to the receiver. If it were not, receivers would stop responding. Consider the phenomenon of aposematism, the signaling to potential predators that a prey organism is distasteful or poisonous. Nature abounds with such signals. For instance, a granular tree frog signals its poisonousness to predators via its distinctive coloration. Here, a sender **S** (the frog) signals **G***y* (I am poisonous) to a receiver **R** (a predator) by **F***x* (the frog's red skin). The kind of information involved in signaling is **C**-information.

As just noted, an organism's signal may cease to be effective if the receiver stops responding. An important element for success, here, is the *reliability* of a signal. Yet false signaling, or "cheating," is a widespread strategy in life. One prominent form of deceit is the phenomenon of Batesian Mimicry. For example, the common Rose Swallowtail butterfly is distasteful to potential predators. It has evolved a unique wing pattern and coloration that presents an aposematic signal. This signal is successful; predators have coevolved to avoid the rose swallowtail. By contrast, the common Mormon Swallowtail butterfly is a perfectly edible prey. But the Mormon Swallowtail has a trick. It has coevolved warning signs that counterfeit the reliable signals of the common Rose. However, this exploit only works because the Rose's signal is reliable. And it turns out that signals are reliable most of the time. This last fact requires explaining.

To ensure the reliability of signals, signalers must pay a cost.

This comes in two forms. First, there is the cost of guaranteeing that the sent **C**-information is adequately received. Think, for example, of having a conversation in a room versus by the roadside. In order to be heard against the din of traffic, the sender must speak louder. This is the cost to be heard.

Second, there is what Maynard-Smith calls the "strategic cost" (Maynard-Smith 2003, p. 17). This is the cost of ensuring that the signal is reliable (i.e., that the sender doesn't cheat). To illustrate, consider the threat displays of two wolves in a contest over some carrion. These displays, such as pricking ears and baring teeth, signal the willingness of a wolf to escalate the contest. Assume that the need of one wolf for food is greater than the other, and that the willingness to fight is less in the wolf that needs food less. If the signals are reliable, the less needy wolf will retreat, the net benefit of the encounter being too low. But why wouldn't a wolf exhibit a threat display whenever convenient, rather than when its need were great? Because of the cost. In this case, the cost of the threat display is the risk of a fight that the wolf might lose.

The terms have finally been set. We can now understand Hutto and Myin's view of NOC: it is the challenge of explaining how the coevolution of signals, consisting only of **C**-information, could give rise to mental states with content. Success, claim Hutto and Myin, hinges on providing a solution to the hard problem of content (HPC).

Section II: The Hard Problem of Content

HPC arises as a problem with a set of claims and commitments that information-based approaches hold. The crucial commitment is to a standard meant to constrain accounts to naturalistic explanations. Hutto and Myin (2013; 2017) call this "explanatory naturalism."

Explanatory naturalism conditions acceptable explanations

of NOC on a commitment to philosophical naturalism. This is a view that is compatible with science in the following sense: Hutto and Myin (2013) define the term by what Wheeler (2005) calls "the Muggle constraint." This is a weak version of philosophical naturalism that is defined by a commitment to ontological physicalism (the thesis that only the physical exists), and the view that philosophy is continuous with the natural sciences. Continuity with natural science is weakly defined as "mere *consistency with* natural science, a reading that makes room, in principle, for multiple modes of explanation" (Wheeler 2005, p. 5). By "multiple modes of explanation," Wheeler means that "natural science" is not restricted to physics but includes social sciences like anthropology, developmental psychology, behavioral economics, and so on. Additionally, Wheeler claims that such continuity implies that if there is a conflict between science and philosophy, philosophy gives way to science. This is the standard to which Hutto and Myin hold both themselves and other information-based approaches.

According to Hutto and Myin (2013), the following set of claims and commitments comprise HPC:

(1) any explanation of NOC is constrained by explanatory naturalism;

(2) accordingly, explanations are limited in terms of **C**-information;

(3) mental content must be explained in terms of **C**-information;

(4) **C**-information is constituted by covariance relations;

(5) mental content consists of **S**-information;

(6) **S**-information is constituted by:

(i) $\boldsymbol{n}$ data (for $\boldsymbol{n} \geq 1$),

(ii) well-formedness,

(iii) meaningfulness; and

(7) **S**-information is neither constituted by nor reducible to **C**-information.

If claims (4)–(6) are true, then (3) cannot be maintained because of claim (7). In other words, "covariance doesn't constitute content" (Hutto and Myin 2013, p. 68). As we saw in Section I.II, claim (8) is true; **C**-information does not meet the three conditions constituting **S**-information. As a consequence of this problem, Hutto and Myin claim that information-based approaches face a dilemma: either retain (1)–(4) but give up (5) and (6), or give up on (1)–(4) and find another naturalistic way to ground mental content. Hutto and Myin take the first option. But as we will see, Hutto and Myin do not reject (5) and (6). Instead, they proceed to redefine the basic nature of intentionality in a way that doesn't require mental content. This move purportedly allows them to carve a path from animal communication based in **C**-information to human communication based in **S**-information. In order to get there, that path must be consistent with sociocultural practices that raise the "question of truth" (Hutto and Myin 2017, p. 119).

Section III: Hutto and Myin's Approach

The approach outlined by Hutto and Myin (2017) combines two phenomena. The first is a contentless form of intentionality they call *Ur-intentionality*, and the second is a form of human niche construction elucidated by Sterelny (2003, 2010) which he calls "scaffolding." This second phenomenon can supposedly lead to "sociocultural practices…that make use of public representational systems in particular ways for particular ends [and account] for both the initial and continued emergence of content-involving minds" (Hutto and Myin 2017, p. 134).

Before explicating both components of Hutto and Myin's approach, I will define the test that I will use to evaluate whether or not they succeed. Any account of the natural origins of content that is limited to **C**-information must show the following:

(1) *Dissociation*. The behavioral interactions mediated by signals (e.g., via vocalization) are instances of covariance of states. These coevolved under the constraints of natural selection and strategic cost. A necessary first step in departing from mere covariance is the dissociation of the *vehicle*, i.e., the *means for* signaling (e.g., vocalization) from their coevolved response(s). Only when an animal has dissociated the vehicle from their coevolved responses can such vehicles then be used as a sign for something else;[7]

(2) *Re-association for reference*. If a vehicle has become dissociated, then it is possible for it to become associated with another thing.[8] *Only if the dissociated vehicle has been re-associated with some other thing can it then be used to refer to that thing;*[9]

(3) *Basic norms for the formation of signs that have propositional form*. In order for a sign to have truth conditions, it must represent the referent as being a certain way. Therefore, a sign has truth conditions only if it has propositional form. For an organism to form a sign into propositional form, it must do so according to some organizing norm(s). An organism need not be able to represent the norm(s), nor must they be conscious of it while using it. *Only if an organism can form a sign into propositional form according to some norm(s), will it have truth conditions*.

Any account attempting to satisfy all three conditions must do so in a way that respects HPC. Since semantic content is regarded, here, as parasitic on mental content, if an account satisfies all three conditions, it will succeed in explaining NOC. That is, in showing how truth-evaluable signs can emerge from the modification of behavior based in **C**-information, such an account succeeds in showing how **S**-information, and thus mental content, naturally originate.

Section 3.1: Ur-Intentionality

Neither *E. coli* nor mosquitoes, frogs, or wolves speak a language. However, as we've already seen, there are natural mechanisms that allow for them to become involved in particular ways with their environment. The kinds of signal-based interactions that coevolve through natural selection and strategic cost allow for basic communication. Hutto and Myin claim that on such a view, a kind of primitive intentionality can be understood as a *directedness* at environmental targets rather than defined as representation. They call this Ur-intentionality.

Ur-intentionality has the following features: (1) it includes a sensitivity to and capability to selectively respond to **C**-information in an environment, (2) it is non-referential, (3) it is directional in the sense that it consists in a coupling of organisms to features of their environments, (4) it can "objectivate," (5) it may be affective (i.e., include emotional states), and (6) it is teleofunctional.

To best capture what Ur-intentionality is supposed to look like, I will start with feature (4). Hutto and Myin are the founders of an approach to explaining cognition that they call "radical enactive cognition" (REC). Like adherents of ecological psychology, they base their view of cognition on the notion of "ecological resonance." Guitars resonate to the sound waves produced by their strings. The cochlea in ears also resonate with the sound waves amplified by the guitar allowing us to hear its tones. The sense of "resonance" in these examples is clear; what "ecological resonance" means is not. To try to make sense of it, let us consider the example of a complex organism. Within the body of an animal, there are several specialized biological systems, i.e., organs formed of tissues made of similar kinds of cells that, as a whole, transform available resources into work. The organs of a body interact in many ways, including forming larger systems. For example, the activity of the blood vessels, arteries, veins, heart, and lungs are coupled together to supply oxygen throughout

the entire body for cellular respiration. This coupling is dynamic in the sense that each system, as well as the system as a whole, is sensitive to changes in any one system, other organ systems of the body (e.g., the digestive system), and the environment.

The idea of resonance applies this notion of dynamic coupling to intentionality. Ur-intentionality "directs" the organism and its behavior towards features of the environment. This is explained in terms of a dynamic coupling of the activity of the central nervous system with some environmental feature. According to REC, this activity isn't pre-programed or limited by specific functions, nor is it passive with respect to the environment. Rather, it is highly flexible and extensive. This latter notion is meant to convey the thesis that the work accomplished by neuronal activity (i.e., "cognition") is partially constituted by systems in the environment to which it is coupled.[10] Moreover, such coupling is dynamic in that the nervous system also affects how it coupled. This active sense of the nervous system's involvement with the environment is what is meant by "enactive." Finally, it is such coupling which constitutes the directionality of Ur-intentionality. That is, constrained by natural selection, an organism's perceptual systems and behavioral repertoire coevolve (i.e., couple) with environmental systems that provide net evolutionary benefits. In this way, they are said to "resonate."

Beyond this vague description of resonance as ecological coupling, the notion remains "theoretically underdeveloped" (Ryan Jr. and Gallagher 2020, p. 2). Firstly, it isn't clear what the kind of directionality portrayed by (5) adds to Maynard-Smith's notions of cue and signal. Those already frame animal interactions as being "involved" with their environments in contentless terms (Godfrey-Smith 2014, p. 81). More to the point, referring to resonance as directedness strains the sense of *intentional* directedness. Resonance doesn't require any "internal extra something," in the sense described in section I (Ryan Jr. and Gallagher 2020, p. 2). Indeed, the claim that REC can explain "a creature's capacity

to perceive, keep track of, and act appropriately with respect to some object or property without positing internal structures," (i.e., mental states) is supposedly one of its virtues (Hutto and Myin 2013, p. 82).

The other six features distinguish Ur-intentionality from other "cognitive" activities. It follows from the description of Ur-intentionality that it isn't referential or propositional. Thus, it doesn't have truth-conditions. It is teleofunctional because its directionality coevolves to accomplish certain jobs that serve the fitness of the organism, and it presupposes that organisms evolve structures to enact sensitivities to **C**-information.

Feature (5) is a term that Hutto and Myin adopt from Roy (2015). Objectivation is the name that Roy gives to a particular interpretation of Franz Brentano's original description of intentionality as a "direction towards an object" (Brentano 1874, p. 67). Roy construes this as a relation to something *as an object*, rather than as a mere thing, where the relatum is objectified by the subject. Further, the directionality of objectivation is immanent to the consciousness of the subject such that they experience being directed at it as an object. Presumably, objectivation is not an essential feature of Ur-intentionality since if it were, we would have to recognize creatures such as mosquitoes as conscious subjects. Perhaps mosquitoes have consciousness. Philosophers disagree about whether they do, anyway.[11]

The question of consciousness aside, it isn't clear what subjectivity could mean in a merely Ur-intentional organism beyond being the locus of a perspective. Any organism that is coupled together with some environmental feature has a unique relationship to that feature, viz., *it's* being coupled to it, as opposed to another organism being coupled to it. But is mere coupling constitutive of a perspective? In one sense, no. If by perspective one means having attitudes about things (e.g., beliefs), then, clearly, resonance falls far short. However, if one means simply having some relationship to particular environmental features,

from a particular location, then, in this weak sense, resonance can include perspective. Yet, even if Ur-intentionality provides for a minimal kind of subjectivity, nothing among its defining characteristics account for how that subject objectivates some feature to which it is coupled. Thus, it isn't clear how objectivation fits into the Ur-intentionality scheme.

With feature (6), Hutto and Myin claim that Ur-intentionality "is compatible with, and embraces" the concept of "expressive behavior" developed in Bar-On and Priselac (2017, p. 143). By Bar-On and Priselac's (2011) lights, expressive behavior opens the possibility that communication based in **C**-information can become something more. Something on the way to linguistic behavior based in **S**-information. In the next section, I will examine the notion of expressive behavior with the aim of assessing whether it adds something to Ur-intentionality that satisfies any of my three conditions for a successful account of NOC.

Section 3.2: Expressive Behavior

According to an oft-cited study by Cheney and Seyfarth (1990), vervet monkeys have three distinct alarm calls for specific predators.[12] In response to large cats such as leopards, vervets produce a loud barking call. Distinct from this, upon seeing either of two kinds of eagle, vervets will produce a short double cough. Finally, when faced with a poisonous snake, a vervet will utter a "chuttering" sound. Each of these signals can be described in terms of **C**-information-based interactions, as I've shown. However, when we observe the behavior of vervets sounding the alarm, or the actions of pets greeting us when we arrive home, our common sense strongly tells us that there's something more to these displays. Bar-On (2013) and Bar-On and Priselac (2011) claim that there is, viz., expressions that "directly reveal" both the psychological states of senders and their causes (e.g., predators or lovable humans) (Bar-On and Priselac 2011, p. 129).

Bar-On and Priselac recognize a wide range of behaviors as falling into the category of expressive behavior. This spans from gasps and grimaces to hat tips to reactive cries of "Ouch!" and finds its limit at speech acts such as commands rich in **S**-information. Yet what distinguishes non-linguistic expressive behavior from linguistic behavior is its constitution as a relation. According to Bar-On and Priselac, non-linguistic expressive behavior isn't a semantic relation between linguistic vehicles and semantic content. Rather it is a psychological relation between the expressive vehicles of signaling behavior (e.g., growling) and states of mind (e.g., agitation). Specifically, non-linguistic expressive behavior reveals the quality (e.g., fear, joy) and intensity of the sender's mental state. What pushes non-linguistic expressive signals past mere covariance of coevolved behavior is its purported "world-centrism" (2011, p. 132).

As an illustration, consider the following scenario involving vervet monkeys. Assume we find two vervets, a sender **S** and receiver **R**, foraging in a woodland. This activity is risky since it exposes **S** and **R** to predation. But as vervets, **S** and **R** have coevolved sensitivities to particularly dangerous features **O** (e.g., leopards). In response to sighting an **O**, the vervets have also coevolved certain kinds of expressive signals **E** (e.g., alarm calls), as well as self-preserving reactions **A** (e.g., fleeing into a tree). Furthermore, as **R**'s, vervets have coevolved certain responses to **E**, namely **A**.

Suppose, now, that **S** spots an **O** and responds with **E**. According to Bar-On and Priselac, in **E**-ing, **S** is showing **R** how they are about to act (viz., **A**). Moreover, **E**-ing is world-centric in the sense that it shows to **R** the mental state of **S** (e.g., fear), its cause **O**, **S**'s anticipated response **A**, and enjoins **R** to **A**. Though **E** does not refer or convey a singular thought, it does direct **R**'s attention to **O**, as well as **S**'s mental state is reaction to **O**. Upon hearing **E**, **R** normally responds with **A**.

The promise for Hutto and Myin in Bar-On and Priselac's

theory of expressive behavior lies in the possibility for dissociation and re-association of a signal's vehicle. To build towards this possibility, consider some additional complexity in our scenario. Let $\mathbf{B}_1$ be **R**'s normal response to **E**; and let $\mathbf{B}_2$ be some response that "disagrees" with the danger signaled by **E**. In normal circumstances, upon hearing **E**, **R** engages in $\mathbf{B}_1$. However, let us suppose that in their foraging, **S** discovers a tasty treat that **R** wants. **S** spots **O** and gives **E**. Now, rather than follow along with $\mathbf{B}_1$, **R**, wanting **S**'s find, responds with $\mathbf{B}_2$. In short, **R** regards **E** as a false signal. Of course, sometimes vervets are wrong about seeing an **O**. And as we know, **E** is subject to strategic cost and organisms may change their orientation to the reliability of such signals. Further, there's always some degree of spontaneity in how animals respond to features of their environments. So, the possibility of **R** engaging in $\mathbf{B}_2$ rather than $\mathbf{B}_1$ is conceivable.

Bar-On and Priselac suggest that the kind of disagreement exhibited by **R** in $\mathbf{B}_2$ opens the door to the possibility of dissociating the vehicles of expressive behavior from the affective states that cause them. And, if dissociation is possible, then re-association is possible, they think. While expressive behavior is intriguing and might be capable of conveying more than just **S**'s mental state, it isn't clear that it gets beyond **C**-information. Expressive behavior is, by definition, non-referential and non-propositional. Bar-On and Priselac's response to this worry is to point out that genuinely linguistic behavior can also be described in terms of **C**-information. If one insists on maintaining that human linguistic behavior should be kept entirely separate from the vervet's non-linguistic expressive behavior, thus implying a discontinuity in the evolution of intentionality, one needs to provide a substantial argument for the separation.

Can Bar-On and Priselac's response secure expressive behavior as a genuine middle stage in the evolution of intentionality, between Ur-intentionality and intentionality proper, for Hutto and Myin? It's hard to see how it can. Firstly, the kind

of framework which construes the vervet's expressive behavior as consisting in mere **C**-information (as in Godfrey-Smith's and Maynard-Smith's signaling framework) is the one to which Hutto and Myin are committed. This follows from their commitment to explanatory naturalism, as well as from their explicit endorsements.[13] That framework is too restrictive to make sense of an intentionality which falls neither into **C**-information-based signals and **S**-information-based signals. Moreover, for this reason, the Bar-On and Priselac response falls far short of resolving HPC. That is, they do not show (beyond hinting at the possibility) that dissociation and re-association have actually taken place in the scenario described above. Thus, something even more is needed to meet conditions (1) and (2). Hutto and Myin claim that that something more is sociocultural "scaffolding."

Section 3.3: "Sociocultural Scaffolding"

Kim Sterelny's (2003, 2010) *scaffolded mind hypothesis* (SMH) suggests that human cognition depends on, and has been transformed by, environmental resources. Moreover, such resources are often maintained, constructed, or altered precisely because they enhance aspects of cognition. To elucidate the import of SMH for Hutto and Myin, I will begin with a short foray into basic ecology. In succession, we'll have a look at the concepts of *ecological niche* and *niche construction*, then return to Sterelny's view that humans alter the **C**-informational character of our niches in ways that transform human cognition.

Organisms such as Sméagol, *E. coli,* mosquitoes, and vervet monkeys all find themselves fitted into *special slots* in their environments, like keys into locks. Like all organisms, they occupy positions in local food-chains that connect, sustain, and moderate their populations. This reflects the fact that, in order to successfully survive and reproduce, lineages of organisms must evolve ways of transforming available local energy sources. Of course, Sméagol

doesn't consume the same resources as mosquitoes do; nor does he acquire his resources by the same methods as vervets. Environments often afford copious and complex sources of energy, potentially supporting wide ranges of diverse life that exploit them in different ways. Those available resources an organism exploits, and the ways they exploit them, define their place or *ecological niche* in the overall environment.

Niches are neither fixed nor are organisms mere patients fitted into them. Organisms also actively alter what their environments are able to afford them in a process called *niche construction*. For example, beavers alter the kinds of resources they may use for shelter by constructing dams. Beavers fell trees, pull them into streams, weight them down with rocks, and reinforce the resulting structure with shrubs and grasses. Completed dams produce ponds and small lakes in which beavers then build aquatic-based shelters protected from predation. In addition, damming alters the local environment by creating new wetlands that afford niches for other organisms like birds and fish.

Another form that niche construction takes is epistemic. Consider the epistemic role of scent marking in defining territory. This behavior modifies the "informational character" of an organism's local environment in cognitively beneficial ways (Sterelny 2003, p. 565). It does so by producing new covariance relations that serve as external memory stores. Rather than an organism remembering the boundaries of its territory, it offloads the epistemic task of navigating safe ground to the informational character (scents) of its environment.[14] In Sterelny's words, it uses its environment to "scaffold" its memory.

Humans are prodigious niche constructors. The most obvious way in which humans modify our environment is through tool use. For example, hand axes and fires clear brush for agriculture. Needles, hair, and hide produce clothing that increase our tolerance to colder climates. Lasers split atoms for the heat to turn turbines. This suggests a long line of learning and innova-

tion. Evidence of tool use and production by hominins such as *Homo erectus*, one of our ancestors, have been found dated as far back as 1.5 million years ago.[15] And unlike other tool users, such as corvids, our technological activities involve a high degree of social cooperation. Like tool use, there's some evidence of social cooperation in *Homo erectus*.[16] Both cooperation and intergenerational learning are deep traits in our evolutionary history and crucial components of SMH.

Sterelny (2010) expands upon this by considering a hypothetical environment for stone tool making. Assuming humans possessed the capacity for imitation, Sterelny reasons that intergenerational learning began through the observation and imitation of parents by children, in the exercise of economic activities, e.g., tool-making. As adults become more dependent on stone tools, production increases, and tool-making sites become littered with the right kind of stones, as well as partial, failed, and completed tools. Compared to their learning conditions, the activity of parents would bias the ways in which their children would engage the environment, providing ample exemplars for trial-and-error learning. As the tool-making process develops, learning becomes increasingly supported by the constructed environment of the production site. Thus, increasingly, the intelligence of new generations becomes scaffolded by progressively enhancing epistemic niches. So the Sterelny story goes.

Hutto and Myin claim that situations like this would lead to mental content. Why might they think this? They're infuriatingly silent. The key factor, they think, is "the question of truth" in the mastery of sociocultural practices such as the one above (Hutto and Myin 2017, p. 134). One way of making sense of this is by entertaining the interaction of a teaching parent and an erring child, in a scenario like the one in the previous paragraph. One assumes the parent would act to correct the child. But how? Signaling based in **C**-information is not capable of representing, so not capable of supporting gesture and speech that could show the child the

mismatch in their error. So, we are left with a situation in which the question of truth has arisen in sociocultural practice, and the need to convey this places pressure on the parent to make tools of their vocalizations, so to speak. What might solicit the latter? Hutto and Myin gesture at other cognitive capacities to fill the gap. Human individuals, they note, are embedded in dynamic, complex environments that change in sometimes random, and often unexpected, ways. Further, their behavior is often spontaneous, owing to traits like curiosity and creativity. If a pre-linguistic hominin such as *Homo erectus* can discover the use of a rock as a knife, why not believe that through play, or perhaps accident, they could discover vehicles of signals for reference or to make a claim, in order to satisfy a social need?

And this is it. No explanation of how the vehicles of signals might become dissociated, re-associated, and take on propositional form in a naturalistically constrained way. But it's worse than that. Recognizing a need to communicate that someone is in error *presupposes* mental content! What is recognizing that someone is in error but having the singular thought that *they are in error*? But Hutto and Myin cannot presuppose content to explain content since not only is this circular but it would fail HPC. On their view, lying underneath cognitive capacities such as curiosity and creativity are processes defined by Ur-intentionality. These are constituted of only **C**-information. So, in the end, Hutto and Myin provide us little more than their fingers pointing at someone else to support the claim that their approach both accounts for NOC and solves HPC. But such finger pointing accomplishes nothing.[17] It neither shows dissociation, re-association, nor the development of norms for forming signs with propositional form. Hutto and Myin's approach fails not only my test, it fails their own.

Section IV: Conclusion

Given everything that has come before, as a reader, you are no

doubt feeling frustrated with this conclusion. Many philosophers, having read Hutto and Myin, feel similarly.[18] In this paper, I have conclusively shown that Hutto and Myin's approach to NOC fails to live up to their claims. I described the standard they claim to meet—solving HPC—and prescribed my own. I articulated the relevant views of philosophers to whom Hutto and Myin outsource their work. And, lastly, I showed that though intriguing in their own right, Hutto and Myin fail to utilize such views to meet either standard for explaining NOC.

Notes

1. See Yamamoto et al. (1990).
2. Omer and Gillies (1971) are cited in Gillies (1980).
3. Maynard-Smith attributes this definition to Hasson (1994).
4. A prominent teleosemantic theory of mental content can be found in Ruth Millikan's *Language, Thought, and Other Biological Categories: New Foundations for Realism*.
5. This is an inadequate description of Teleofunctionalism. Unfortunately, given constraints of space, it will have to do.
6. Godfrey-Smith (2014) calls this the "signal cost."
7. My use of "sign" is synonymous with other philosophers' and linguists' use of "symbol," i.e., a vehicle which represents some other thing.
8. Association is here meant only to indicate that the reference relation has been established. I do not take any stand on a theory of reference fixing.
9. An *intention* understood as a singular thought with content, e.g., *I will dissociate this grunt from its behavior and re-associate it with that rock*, is not presumed. Such a presumption would be circular. Rather, dissociation and re-association is only assumed to be constrained by a hominid's current cognitive abilities.
10. I wrap "cognition" in quotation marks, here, since their view of cognition is starkly different from that of most cognitive scientists.
11. Of course, belief that mosquitoes are conscious depends *inter alia* on one's theory of consciousness.
12. Cited from Maynard-Smith (2003), pp. 113-114.
13. Godfrey-Smith and Maynard-Smith base their frameworks on evolutionary game theory and evolutionary biology. Godfrey-Smith adopts Lewis' (1969) signaling model (see p. 80). Maynard-Smith assumes game theory (see 2003,

chapters 3-5). Hutto and Satne (2015) explicitly endorse Godfrey-Smith's view on p. 531.

14. The sense of "epistemic" is fairly weak here. It conveys that organisms are guided by **C**-information in their environments.
15. For instance, see Semaw, et al. (2020).
16. For instance, see Hatala, et al. (2016).
17. It is worth noting that Sterelny doesn't think SMH can yet provide an account of NOC, saying in his (2003) that "while the argument of this book does not refute [the] naturalization project, it does not vindicate it" (p. 879). Indeed, for him, language seems to be already present in the kind of sociocultural practices entertained above. It is, for him, one of its "fuels for success."
18. See for instance, Shapiro (2014), Rowlands (2015), Roy (2015), or Rupert (2018).

Bibliography

Bar-On, Dorit. (2013) "Expressive Communication and Continuity Skepticism," *Journal of Philosophy* 110(6), pp. 293-330

Bar-On, Dorit and Priselac, Mathew. (2011) "Triangulation and the Beasts," C. Amoretti and G. Preyer (Eds), *Triangulation. From an Epistemological Point of View*, pp. 121-152 (Heusenstamm, Germany: Ontos Verlag)

Brentano, F. C. (1874/2009) *Psychology from an Empirical Standpoint* (New York, NY: Routledge & Kegan Paul)

Campos, M.V., and Gutiérrez, Antonio M.L.. (2015) "The Notion of Point of View," in: Campos and Gutiérrez (Eds), *Temporal Points of View: Subjective and Objective Aspects*, pp. 1-58 (Cham, Switzerland: Springer International Publishing)

Cohen, Johnathan. (2001) *Information and Content* (website): https://aardvark.ucsd.edu/mind/informational_semantics.html

Crane, Tim and Azzouni, Jody. (2011) "Singular Thought," *Aristotelian Society Supplementary Volume* 85(1), pp. 21-43

Floridi, Luciano. (2010) *Information: A Very Short Introduction* (New York, NY: Oxford University Press)

Gillies, M.T. (1980) "The Role of Carbon Dioxide in Host-Finding by Mosquitoes (Diptera: Culicidae): A Review," *Bulletin of Entomological Research* 70(4), pp. 525-532

Godfrey-Smith, Peter. (2007) "Information in Biology," in: Hull and Ruse (Eds), *The Cambridge Companion to the Philosophy of Biology,* pp. 103-119 (Cambridge: Cambridge University Press)

_____ (2014) “Signs and Symbolic Behavior,” *Biological Theory* 9(1), pp. 78-88

Hatala, Kevin G., and et al. (2015) “Footprints Reveal Direct Evidence of Group Behavior and Locomotion in *Homo erectus*,” *Scientific Reports* 6(1) pp. 1-9

Hutto, Daniel D. and Myin, Erik. (2013) *Radicalizing Enactivism: Basic Minds without Content* (Cambridge, MA: The MIT Press)

_____ (2017) *Evolving Enactivism: Basic Minds Meet Content* (Cambridge, MA: The MIT Press)

_____ (2018) “Going Radical,” in: Newen, de Bruin, and Gallagher (Eds), *The Oxford Handbook of 4E Cognition*, pp. 95-115 (Oxford: Oxford University Press)

Hutto, Daniel D. and Satne, Glenda. (2015) “The Natural Origins of Content,” *Philosophia* 43(3), pp. 521-536

Laidre, M. E., and Johnstone, R. A. (2013) “Animal signals,” *Current Biology* 23(18), pp. R829–R833

Lycan, William G. and Neander, Karen. (2008) “Teleofunctionalism,” *Scholarpedia* 3(7): 5358

Maynard-Smith, John and Harper, David. (2004) *Animal Signs* (New York, NY: Oxford University Press)

Mech, L. David. (1970) *The Wolf: The Ecology and Behavior Of An Endangered Species* (New York, NY: The Natural History Press)

Pocheville, Arnaud. (2015) “The Ecological Niche: History and Recent Controversies,” T. Heams, P. Huneman, G. Lecointre, and M. Silberstein (Eds), *Handbook of Evolutionary Thinking in the Sciences*, pp. 547-586 (Dordrecht: Springer)

Ryan, KJ Jr., and Gallagher, Shaun. (2020) “Between Ecological Psychology and Enactivism: Is There Resonance?” *Frontiers in Psychology* 11, pp. 1–13

Roy, Jean Michel. (2015) “Anti-Cartesianism and Anti-Brentanism: The Problem of Anti-Representationalist Intentionalism,” *The Southern Journal of Philosophy* 53(S1), pp. 90-125

Semaw, S., and et al. (2020) “Co-occurrence of Acheulian and Oldowan Artifacts with *Homo-erectus* cranial fossils from Gona, Afar, Ethopia,” *Science Advances* 6(10), pp. 1-8

Sterelny, Kim. (2203) *Thought in a Hostile World: The Evolution of Human Cognition* (Hoboken, NJ: Blackwell Publishing Ltd.)

_____ (2010) “Minds: Extended or Scaffolded?” *Phenomenology and the Cognitive Sciences* 9(4), pp. 465-481

Stiner, Mary C., and Kuhn, Steven. (2006) “What’s a Mother to Do? The Division of Labor among Neanderthals and Modern Humans in Eurasia,” *Current Anthropology* 47(6), pp. 953–980

Wadhams, George H., and Armitage, Judith P. (2004) “Making Sense of It All: Bacterial Chemotaxis,” *Nature Reviews Molecular Cell Biology* 5(12), pp. 1024–1037

Wheeler, Michael. (2005) *Reconstructing the Cognitive World: The Next Step* (Cambridge, MA:The MIT Press)

Yamamoto, Kimiko, and Macnab, Robert M., and Imae, Yasuo. (1990) “Repellant Response Functions of the Trg and Tap Chemoreceptors of *Escherichia coli*,” *Journal of Bacteriology* 172(1), pp. 383-388

The Idols of the Mind in Modern American Political Economy

Sabrina Anne Pirzada

Introduction

In *The Rational Public*, Benjamin I. Page and Robert Y. Shapiro (2005) argue that our policy preferences as a collective are rational. This is not a new argument. In philosophy and social science, we call this "the wisdom of the crowd." However, recent studies have shown that the wisdom of the crowd effect can be damaged by social influences. In this paper, I will show through the lens of Francis Bacon's (2019) "idols of the mind," how cognitive biases can lead to a diminishing impact on our ability to behave in rational ways as a collective. This argument is intended to provide hope that in recognizing the critical flaws in our humanity, we can attempt to become better versions of ourselves and build more rational economic policies.

In Section I, I will explore the wisdom of the crowd effect, what it means to be rational, and arrive at the conclusion that the wisdom of the crowd objective may be injured or impaired when cognitive errors are exercised by large numbers of people. Throughout this paper, I will illustrate this using Bacon's "idols of the mind" concept and provide modern economic examples of how these cognitive errors can get in our way if we are not cognizant of them. In Section II, I will give an overview of the idols. In Sections III through VI, I will detail the way in which each of these idols work. By the end of this paper, I hope to have found value in adapting this old idea into an application that points toward solutions for some of our modern problems.

Section I: The Wisdom of the Crowd

For thousands of years, the wisdom of the crowd has been a phenomenon demonstrating that we as a people are capable of more together than we are apart. Historically, Aristotle (1944) was the first person we know of to propose this notion: "it is possible that the many, though not individually good men, yet when they come together may be better, not individually but collectively, than those who are so, just as public dinners to which many contribute are better than those supplied at one man's cost" (Aristotle 1944, Pol. 3.1281b). A way that we can account for this mathematically for example, is demonstrated through the Law of Large Numbers. This law essentially states that when an experiment is run a sufficiently large number of times, the greater our chances that the error will eventually shrink to zero. This is a mysterious and exciting phenomenon because, in theory, the more people who vote, the greater our chances of getting the most representative answer. We know that while there are implicit assumptions and noise associated with individual judgements, extracting the average over a large diverse body of responses will achieve some degree of success in cancelling out a lot of this noise.[1] To me, it's not clear that we've always gotten it right in the past and I suspect that mass quantities of cognitive biases have at least something to do with this.

Alexis DeToqueville (1966) talked about "the tyranny of the majority." We know from years of debate about utilitarianism that sometimes what is best for the greatest number of people can often causes a remarkable amount of pain for a minority of people. In economics, we call this a "Zero Sum Game," meaning there will always be some tradeoff or level of inequality in order to benefit a group of economic actors. In order for someone to win, someone has to lose. The "Pareto Optimality" occurs when we achieve a system in which no one loses. Most economists tend to view this as a pipe dream, but that does not mean we should not strive for it.

While taking the average is an excellent method for eliminating random errors, it is inadequate for eliminating systematic errors that sway the preferences of the crowd in similar ways,[2] which should concern us about how rationality is measured in our society today. British statistician George Box once noted, "All models are wrong. Some are useful."[3] While the wisdom of the crowd effect irons out random errors over time, it cannot be expected to offset cognitive biases of the collective. While it is true that scientific research has proven social groups to be remarkably and mysteriously cognizant when their averaged decisions are compared to decisions of individuals, it is alarmingly evident that the wisdom of the crowd objective is easily subverted by way of social influences and cognitive errors exercised by large numbers of people. These problems are worth our attention.

Rationality is our primary epistemological vehicle for testing whether or not something is true. We operate through intellect and deductive reasoning and have faith in only what we can see, touch, and prove. We also operate under the truth that one of the things we know is that we do not know everything. As Alexander Pope (2013) said in his *Essay on Man*, "So man, who here seems principal alone / Perhaps acts second to some sphere unknown / Touches some wheel, or verges some goal / 'Tis but a part we see, and not a whole" (Pope 2013, p. 1247).[4] In order to be a truly ethical scientist, mathematician, economist, or philosopher one must operate under the assumption that we do not know everything, which requires both a humble curiosity and acceptance of the paradoxical nature reality where many principles are certain, many are uncertain and imprecise.

Section II: The Idols of the Mind

In the early part of the seventeenth century, Francis Bacon published *The Novum Organum* which translates to "The New Organon."[5] The title is a nod to Aristotle's *Organon*, a philosoph-

ical treatise on logic. In this work, Bacon explores cognitive errors in human rationale he refers to as the "idols of the mind" which point to a distressing possibility of collective madness if ignored. If individuals are not careful in their relations with each other and how those relations lead to opinions that translate into laws, even the most powerful organized empires can fall.

For Bacon, there are four idols of the mind that prevent us from arriving at logical conclusions. These are *Idola tribus* (idols of the tribe), Idola specus (idols of the cave), *Idola fori* (idols of the marketplace), and *Idola theatri* (idols of the theater). Here the word "idol" signifies a distraction that causes us to keep ourselves from the truth (on analogy with how the word is used in Christianity). The idols of the tribe occur when "man's sense is falsely asserted to be the standard of things (Bacon, 2019 p.8)." In Bacon's time, one of these falsehoods would have been the belief that the earth was the center of the universe. The idols of the cave are the "idols of the individual man," which make individuals closed-minded or unreceptive to rational ideas due to biases that come from their individual upbringing, education (or lack thereof), environment, and experiences. It's similar to the idea that each is a prisoner of their own time. The idols of the marketplace are "from a bad and unapt formation of words" that provides a "wonderful obstruction to the mind (Bacon 2019, p.8)." These idols have to do with the fact that our words carry associations and the ways in which we can be easily manipulated and misled through language. According to Bacon, these idols were the most concerning. The idols of the theater are the fallacies that we receive through different dogmas or schools of thought that cause us to blindly accept an ideology regardless of whether or not all parts of it make sense. If you are human, you have idols. No one is exempt and there is no amount of magic happening within the bounds of a mathematical model that can fully erase them.

Section III: The Idols of the Tribe

The idols of the tribe are the inherent characteristics of the perceptions within us that behave as false mirrors. It is in our human nature to often choose unsound conclusions. It is also plausible that these false mirrors are prevalent actors in themselves in the way we organize ourselves and construct our view of the world around us. Bacon states that these idols arise from the constitution of our essence in "limited faculties or restless agitation" or from "the inference of passions or the incompetence of the senses." Similarly, Alexis DeToqueville stated in *Democracy in America* that "experience plunges mankind into universal doubt and distrust" (DeToqueville 1966, p. 187). In *Sapiens*, Historian Yuval Noah Harari (2015) provides an account of human history and explanations for how societies and our sensibilities have taken shape over time. For example, because social cooperation is paramount for our species to survive and reproduce, we had to learn how to gossip. This was a survival technique we learned in order to communicate with one other about who and what to stay away from and who to trust. This began dividing us into different bands over time. However, the "truly unique feature of our language is not about the ability to transmit information about men and lions. Rather, it's the ability to transmit information about things that do not exist at all. As far as we know, only Sapiens can talk about entire kinds of entities that they have never seen, touched, or smelled" (Harari 2015, pp. 22-24). The idols of the tribe when applied to democracy at scale can result in a deep and misguided trust or distrust of others in the world economy. According to the Pew Research Center, as of 2016, 49% of Americans believed that U.S. involvement in the global economy is a bad thing and 7% did not know what to believe (Pew Research Center, 2020). We know that some countries in the world are wealthier than others and have comfortable lives while also spending less time working than others. In *The Wealth of Nations*, Adam Smith (2003)—

widely regarded as the father of capitalism—argued that what gives us this opportunity is free trade. The United States is the world's largest economy and therefore, comes from a position of privilege. Throughout our history, we've been at an advantage to make rules that work for us. Yet, we remain very wary and distrustful of others, which can cause us to be inconsistent and unstable in our intentions.

Section IV: The Idols of the Cave

The idols of the cave are the biases of the individual that "intercepts and corrupts the light of nature, either from his own peculiar and singular disposition, or from his education… and the authority acquired by those whom he reverences and admires, or from the different impressions produced on the mind" (Bacon 2019, p. 8). These cognitive biases include the most heinous ones like racism and sexism, which are especially harmful to societies. According to FBI hate crime statistics, in 2016 the number of assaults against Muslims in the United States surpassed the number of assaults that took place in 2001, the year of the September 11 terrorist attacks. According to the Pew Research Center, 75% of American adults who are Muslim find that there is "a lot" of discrimination against this group and nearly 7/10 Americans in the general public agree with them (Kishi 2020).

There is likely a correlation between these numbers and the amount of negative commentary and threats of Muslim registries and travel bans that have been made by elected officials in recent years. It exemplifies how dangerous the idols of the cave are, especially when exhibited by those in powerful leadership roles. The irony in this is that Muslims are more likely than the general public to believe that targeting civilians for political causes can never be justified. Studies show that 84% of Muslims believe that violent tactics against civilians can rarely or never be justified in order to further a political agenda compared to only 59% of the

general public who believe violence against civilians can never be justified (Kishi 2020). When Francis Bacon was conceptualizing the idols of the cave, it seems likely that he may have been thinking of Plato's allegory of the cave. Here biases and lack of education are illustrated in the idea of prisoners in a cave who are unable to see a fire and the puppeteers behind them. What the prisoners think they see it really a shadow cast by something they do not actually see at all. Only upon release from their chains can they turn to see the truth.

Section V: The Idols of the Marketplace

The idols of the marketplace are formed through reciprocal dialogue between people engaged in commerce and social exchange with one another. Bacon claims that an inept arrangement of words will pose a "wonderful obstruction to the mind." Being learned does not grant us complete immunity from this at all times. We must concede that our words force belief and perception that may lead to confusion. Because of this we are often subject to manipulation, in particular through words used by those in positions of influence.

A more modern conception of flaws in our language was presented by American psychologist Marshall Rosenberg (2015) in his theory of *Nonviolent Communication* that there are ways in which we communicate with one another that are "life-alienating," i.e., that prevent us from exhibiting reason (and experiencing compassion) in our relations with one another. He described four classifications of "life-alienating communication." These are (i) *moralistic judgements* (i.e., who is what); (ii) *making comparisons;* (iii) *denial of responsibility;* and (iv) *communicating our desires as demands* (i.e., thinking based on who deserves what). For Rosenberg, this type of communication both derives from and reinforces hierarchical societies where "large populations are controlled by a small number of individuals to those individuals'

own benefit. It would be in the interest of kings, czars, nobles and so forth that the masses be educated in a way that renders them slavelike in mentality" and "the more people are trained to think in terms of moralistic judgements that imply wrongness or badness, the more they are being trained to look outside themselves—to outside authorities—for the definition of what constitutes right, wrong, good, and bad. When we are in contact with our own feelings and needs, we no longer make good slaves or underlings" (Rosenberg 2015, pp. 15-24). The idols of the marketplace are prevalent in society today and they cause us to be manipulated by those in power.

In 1927, The Federal Radio Commission[6] became subject to the Radio Act which established broadcast licenses that were centered around the interest of the public. In 1949, the Federal Communications Commission (FCC) was established and held that broadcasters were required to present both sides of an issue in an honest and equitable manner. This was known as "the Fairness Doctrine." It was revoked in 1987. This meant that broadcasters were no longer required to report on both sides of an issue and therein marked the beginning of a transition for many outlets from a mode of objective journalism to subjective entertainment. Individuals then began to gravitate toward outlets according to their preferences and biases. By way of advances in computer science and machine learning, our devices and social platforms today automatically learn our preferences, thrusting each person into their own personalized echo chamber. Studies show that the U.S. today is exceptionally polarized. A study conducted in 2020 found that an overwhelming majority (~90%) believed that "lasting harm" would occur if the opposing party's candidate were to win the election (Dimock and Wike 2020). The idols of the marketplace are at work in our society today and this should worry us: polarization has measurable impacts on the wisdom of the crowd effect from being realized in a way that is optimal. In a study done at the Swiss Federal Institute of Technology, it was shown that social

influences do indeed undermine the wisdom of the crowd effect as these influences lower diversity of the crowd without correcting the collective error: "our results underpin the value of collecting individuals' estimates in the absence of social influence. However, in democratic societies, it is difficult to accomplish such a collection of independent estimates, because the loss of diversity in estimates appears to be a necessary byproduct of transparent decision-making processes. For example, opinion polls and the mass media largely promote information feedback and therefore trigger convergence of how we judge the facts. The wisdom of crowd effect is valuable for society, but using it multiple times creates collective overconfidence in possibly false beliefs" (Lorenz et al 2011, p. 9024). It is in this way that the idols of the marketplace quite literally impact our ability to make rational choices as a collective, suppressing the wisdom of the crowd.

Section VI: The Idols of the Theater

The idols of the theater cause us to disregard reason in order to participate in a group or a team. A modern example of this is embodied in the two-party system we have in the United States today. In this we plug into an ideology or belief system and accept that everything about it is true even when aspects of it are contradictory. An interesting and timely example of this kind of thinking today occurs when we think about what "conservatism" means in America. To be conservative, by definition of the word, means "to conserve" or to "keep things the way they are." However, when the very same principle is applied to land management or ocean or atmospheric conservation, American conservatives seem to totally lack this preference. The concept of the idols of the theater is also related to the idea that in the United States, we are not just red states and blue states, but our cultural divisions run even deeper. In Colin Woodard's (2011) *American Nations*, it is asserted that the U.S. has been "deeply divided since the days of Jamestown

and Plymouth" (Woodard 2011, p. 2). Colonies were settled by people from disparate backgrounds from Britain, France, Spain, and the Netherlands who had diverse political, religious, and cultural characteristics. They saw each other as the competition and vied for land and capital. Woodard argues that these disparate cultural identities are still embedded in our social fabric and that America has never been simply one idea, but rather several different versions of an idea. Because of the way in which the idols of the theater are alive and at work today, we often behave as rival nations within the bounds of the same overarching system of government in spite of the fact that we carry the same name, the same constitution, and are seen on the world stage as a singular body. Because we are essentially divided up into teams that each group a complex set of ideological beliefs into one basket, the idols of the theater impact our ability to come to rational conclusions on a number of issues. A study conducted in 2020 found that 24% of the population who backed the Republican nominee were concerned about the coronavirus outbreak and 84% were concerned about the economy. Meanwhile, over 82% of the population who backed the Democratic nominee were concerned with the coronavirus outbreak and 66% with the economy. If the idols of the theater were not at work, these numbers would have a higher degree of equality to them as both the safety of our citizens and a healthy economy for us to survive in are issues to be profoundly concerned about.

Conclusion

It is clear that the idols of the mind are alive and at work today. When cognitive errors of these kind are exhibited in large numbers, it inhibits our ability to make rational decisions as a society and subverts the wisdom of the crowd effect. The idols of the tribe cause us to be distrustful on a global scale; the idols of the cave cause us to discriminate against ethnic and religious minorities within our

own society and to therefore, turn inward and inflict violence on ourselves; the idols of the marketplace polarize us therein keeping us from being rational and productive; and the idols of the theater cause us to blindly reinforce this divide. Because of these errors, we have not stepped into the fullness of what we as a society are capable of in terms of rationality and wisdom. The greater effort we put forth in reducing these cognitive errors, the greater our chances of becoming a society that expresses itself in accordance with the wonder that is wisdom.[7]

Notes

1. It should be noted that when we apply this to democracy, it emphasizes the paramount importance that as many diversely and divergently opinionated, eligible citizens as possible in a given society exercise voting.
2. Polarization is a very timely, important example of this I will address in section V.
3. I would like to thank my friend Thomas Gibson who made me aware of this quote.
4. I owe a debt of gratitude to Dr. Matthew Brosamer at Mount Saint Mary's University in Los Angeles, CA for instilling in me an enthusiasm for this kind of literature and for reviewing this paper.
5. Organon means "instrument."
6. The Federal Radio Commission is the predecessor of the FCC.
7. I am grateful to Professor Mark Balaguer in the Department of Philosophy and Professor Sandor Ferencz in the Department of Economics at California State University, Los Angeles for reviewing this paper.

Bibliography

Abdo, Geneive. (2020) "U.S. Muslims Are Concerned about Extremism in Name of Islam," *Pew Research Center*, Pew Research Center, 30 May 2020, www.pewresearch.org/fact-tank/2017/08/14/like-most-americans-u-s-muslims-concerned-about-extremism-in-the- name-of-islam/.

Aristotle. (1944) *Aristotle in 23 Volumes*, Vol. 21, H. Rackham (Tr) (Cambridge, MA:Harvard University Press)

Bacon, Francis. (2019) *Novum Organum*, pp. 5–70 (Whithorn, UK: Anodos Books) DeTocqueville, Alexis. (1966) *Democracy in America*. J.P. Mayer (ED, George Lawrence (TR) (New York, NY: Harper & Row)

Harari, Yuval Noah. (2015) *Sapiens* (New York, NY: Harper Collins)

Dimock, Michael, and Richard Wike. (2020) "America Is Exceptional in the Nature of Its Political Divide." *Pew Research Center*, Pew Research Center, 13 Nov. 2020, www.pewresearch.org/fact-tank/2020/11/13/america-is-exceptional-in-the-nature-of-its-political-divide/.

Kishi, Katayoun. (2020) "Assaults against Muslims in U.S. Surpass 2001 Level." *Pew Research Center*, Pew Research Center, 30 May 2020, www.pewresearch.org/fact-tank/2017/11/15/assaults-against-muslims-in-u-s-surpass-2001-level/.

Lorenz, J., et al. (2011) "How Social Influence Can Undermine the Wisdom of Crowd Effect." *Proceedings of the National Academy of Sciences,* 108:22, pp. 9020–9025.

Matthews, Dylan. (2011) "Everything You Need to Know about the Fairness Doctrine in One Post," *The Washington Post*, 23 Aug. 2011, www.washingtonpost.com/blogs/ezra-klein/post/everything-you-need-to-know-about-the-fairness-doctrine-in-one-post/2011/08/23/gIQAN8CXZJ_blog.html.

Page, Benjamin I., and Robert Y. Shapiro. (2005) *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences* (Chicago, IL: University of Chicago Press)

Plato, *Republic*, VII 514 a, 2–517 a7

Pope, Alexander. (2013) *Essay on Man*, Tom Jones (ED), *The Norton Anthology of English literature: The major authors*, pp. 1247 (New York: W.W. Norton & Company)

Pew Research Center. (2020) "Widespread Uncertainty About America's Place in the World." *Pew Research Center - U.S. Politics & Policy*, Pew Research Center, 28 July 2020, www.pewresearch.org/politics/2015/05/05/public-uncertain-divided-over-americas-place-in-the-world/.

Rosenberg, Marshall B. (2015) *Nonviolent Communication: a Language of Life*. (Encinitas, CA: Puddledancer Press)

Smith, Adam. (2003) *The Wealth of Nations* (New York, NY: Bantam Dell)

Woodard, Colin. (2011) *American Nations*. (London: Penguin Books)

Doing Our Duty: the Moral Impermissibility of Suicide

Hudson Olander

Introduction

Over the course of human history, few topics have been more controversial than suicide. In Western culture, the act of suicide has been vilified for various social, religious, and moral reasons. However, just as often as not, it seems that society desperately tries to justify its revulsion to suicide without providing a well-reasoned account of why, precisely, it is immoral. While historical-cultural attitudes have drifted, almost aimlessly, between complete condemnation and surreptitious glorification, the modern, educated, Western perspective—which has often been influenced by contemporary scientific, anthropological, psychiatric, and psychological trends—has been one of countenance under reasonable circumstances.

Whether this modern view is a reaction to the almost superstitious beliefs about suicide of the past or an adaptation of libertarian ideals of personal freedom, is beyond the scope of this paper. Rather than focusing on why people choose to adopt certain views on the moral permissibility of suicide, in this paper I will be addressing the four major moral models to which appeals are commonly made in moral claim regarding suicides. I will describe the only reasonable ethical approach to the complex interplay of moral responsibilities that suicide entails. The model's ultimate conclusion will show that suicide is almost never morally permissible. Following this, I will describe some common intuitions that suggest suicide is morally permissible, and I will show why these intuitions are flawed. Finally, I will show that, even when suicide

would seem permissible under the reasonable moral model I present, there are still problems with identifying such scenarios as permissible.

Section I: The Moral Models Worth Consideration

With any discussion regarding suicide, the first step must always be to explicate the term itself. At first blush it seems natural to assume that any instance in which someone is causally responsible for their own destruction could be reasonably described as suicide. An in-depth examination of specific instances of self-destruction, however, often reveals that the metaphorical waters are quite murky. Indeed, it is often the case that one's intuitive judgments on the moral nature of suicide color one's perceptions regarding the nature of a self-inflicted death (Cholbi 2011). Not every case is as cut-and-dried as a person putting a noose around their neck and jumping off of a chair. Imagine a man who mistakenly takes a night of heavy drinking to the extreme and dies of alcohol poisoning; did this man commit suicide? A soldier sees a live grenade tossed in front of her comrades and leaps on it to shield them from the explosion. Was her heroism a suicidal act? An elderly person spends sixty years of their life smoking cigarettes, despite knowing the health risks involved, and ends up dying of lung cancer. Did they choose to end their life? What of people who request a doctor's help in acquiring drugs for euthanasia? What of the person who takes a bite of food that happens to be poisoned, resulting in their death?

With these questions in mind, we can assert that there must be necessary components for a self-caused death to be considered a suicide. An act can be thought of as suicidal if someone is the cause of their own death, and they caused their death intentionally rather than through coincidence (Fairbairn 1995). If someone decided to go to the bank, coincidentally on the day a bank robbery

occurs, and they get shot to death in cross fire between police and robbers, one wouldn't reasonably consider the act of going to the bank that day a suicidal one. Therefore, for an act to be considered a suicide, one must be causally responsible for the chain of events that leads to their death, and they must intentionally be involved in said causal chain. Thus, while a smoker who dies of lung cancer is causally responsible for the chain of events that led to their death, they have not committed suicide because they didn't participate in the causal chain with the intended outcome being their death. Conversely, a patient's requesting for euthanasia is an act of suicide—while they don't administer their own euthanasia drugs, they are part of the causal chain of the administration because of their request to their doctor and the intention of their actions is their destruction.

Under this model, it is the intention rather than desire per se that matters. While perhaps controversial, this means that even seemingly noble acts of martyrdom are acts of suicide. The soldier jumping on a grenade, whose actions may intuitively seem permissible, is an act of suicide. One might argue that the soldier's act is not suicidal because their motivation is to save others, rather than to die. If one throws themselves upon a powerful explosive, however, grievous bodily harm or death are reasonably expected outcomes. The soldier may desire the safety of their fellows, but they intend their actions to end result their own destruction.

Now that we have a working model for determining what sorts of actions can be considered suicide, we can begin to evaluate the most common moral models that address the ethical status of the act. These models include a theological model, a utilitarian model, a libertarian model, and a deontological model. While the first three models differ in their reasoning and conclusions, I will show that their claims are either highly improbable, or sow the seeds of their own defeat. Only the deontological model possesses the tools necessary for giving a complete and consistent account of the ethical weight of suicide.

The first moral account of suicide that needs to be considered is one in which has heavily influenced Western cultural attitudes since medieval times: Christian theological philosophy (Amundsen 1989). By expanding on arguments first outlined by St. Augustine, Thomas Aquinas gave what is commonly thought of as the origin of modern Christian philosophy's prohibition against suicide. Aquinas lays out three points: first, suicide is incompatible with the natural self-love that God instilled in humans to facilitate their preservation; second, suicide damages the subject's community; and finally, because God gave life to humans as a gift, ending that life purposefully violates humanity's duty to respect the gifts of God. God has the right to determine how long each person's life shall last and ending it prematurely is essentially usurping that right (Aquinas 1271). With this claim, Aquinas concludes that suicide must be morally impermissible.

All three of Aquinas' arguments depend on multiple unsupported assumptions. The most obvious of these is the existence of God. Without reliable proof of the existence of God, none of Aquinas' points carry any argumentative weight. There can be no incompatibility with that which God has instilled in humans if there isn't a God to have done any instilling; if there is no God, then we cannot act inconsistently with His wishes. Likewise, with Aquinas' third argument, ending one's life certainly can't be a violation of one's duty to God only if God exists. Even his second, and most seemingly secular argument, relies on the understanding that being an asset to one's community is important because it is in service to God. Without the acceptance of God, it is not readily apparent that one has the kind of duty to their community that Aquinas presupposes.

Even if God exists, Aquinas' argument still faces serious problems as it also requires acceptance of a very particular and somewhat baffling metaphysics. One might be willing to accept God's existence without also accepting that suicide ends a life before its preordained time. One might suppose that God decided

a woman would die by suicide at the age of thirty-two, and that if she fails to end her life, she is usurping God's right to decide the time of her death. A very strange version of compatibilism must be accepted to believe in an omniscient God whose plan for the future can be defied by human action. Likewise, we must accept that God wishes all humans to contribute directly to their communities. This seems to run contrary to the Christian tradition's acceptance of the practice of ascetic monks who live alone as divine hermits.

While Christian theological philosophy fails to give a plausible account for the moral impermissibility of suicide for those who do not already accept Aquinas' metaphysical assumptions, utilitarianism attempts to present a moral model that carries no such additional baggage. The utilitarian argument for the moral permissibility of suicide is that a sort of moral weighing needs to be performed for any action. The utility that would arrive from a person's continued life must be compared to the utility that will be destroyed by the person's death. If we imagine a person feeling tremendous anguish, wishing to end their life, then the utilitarian perspective recommends that a calculation must be made regarding how much good, or utility, the person can still bring into the world in their anguished state and how much suffering could be averted by their suicide.

It is relatively non-controversial to claim that suicide harms the friends and family of the person who performs the suicide. The emotional pain of loss, the outrage, the extreme feelings of loneliness, and the vulnerability engendered by a loved one's suicide all constitute negative utility in themselves. Additionally, they also can create a drop in productivity, thus compounding the negative utility. Some might claim that this negative utility is created because of the cultural perceptions of suicide, rather than the suicide itself (Pabst Battin 1996)—that if suicide was approached in a different way by society, then those around a person who commits suicide would be able to accept the now-

deceased person's actions, and consequently wouldn't suffer from psychological turmoil. While this position can counter the claim that a suicide always results in negative utility to society, it fails to account for the utility that is potentially lost because a person who ends their life can no longer generate utility. Even someone who suffers from great mental or physical anguish has some capacity to generate utility. Regardless, this view still requires a way to account for how much overall utility is produced or averted by the suicide.

It is here that the question of the permissibility of suicide encounters utilitarianism's most obvious problem: how does one measure utility? What sort of units can account for something as broad and abstract as utility? Furthermore, how can that sort of measurement account for negative utility? Even if one were to decide on some sort of unit for utility, what sort of apparatus would one deploy to perform the calculation? Conceiving of a device that can measure the producible utility of a man who is in constant pain, against the grief his widow faces would require a buy-in almost on the scale of Aquinas' metaphysics. While the utilitarian moral model implies that suicide can be permissible or impermissible on a case-by-case basis, it provides no practical way to determine which cases are which.

While the search for a means of utilitarian measurement may be an intractable problem, the libertarian view offers an almost inversely simple account of the moral permissibility of suicide. Under the libertarian moral model, suicide is morally permissible because of a person's right to end their own life. This right is typically outlined in one of several different ways. The first is that the right to commit suicide is a type of right of noninterference (Pabst Battin 1996). Those who take this view assert that others do not have the right to interfere with any individual's freely made choices, so long as those choices do no harm to anyone else. This could also be interpreted as a libertarian liberty right, which claims an individual always has the freedom to act in whichever

way they see fit as long as no one else is harmed by those actions.

These rights, according to the libertarian view, are derived from the ownership of our bodies. Imagine I own a car, I could be reasonably thought to be morally permitted to do whatever I so choose with it, provided I do no harm to anyone else. On this view, if a person owns their body, then they have to right to treat it however they wish and they should have the right to make those decisions without significant interference from others. We intuitively believe that the owner of a toaster may use it however they like, even to the extent of destroying the toaster. This ownership one has of one's own body runs counter to the Christian theological notion that God has ownership of our bodies.

The most common concern with this ideological line is that a person doesn't own their body in the way that they might own a car or a toaster (Cholbi 2011). This claim revolves around the notion that to own something there must be a metaphysical distinction between the owner and the thing which is owned. One doesn't own one's body, because a body is part of what it means to *be* a particular person. We can show that this is more than a mere syntactic distinction by contrasting the sorts of actions humans can take with owned objects from the sorts of actions one can take with their body. For example, one can give away their toaster, but one cannot give away their body.

While the exact nature of the relationship that one has with one's body is debatable, libertarianism fails to accurately capture the moral weight of suicide in another manner. Suicide is almost always, if not always, a source of harm for others. The libertarian rights of noninterference and liberty hinge on the notion that the sorts of actions the free individual undertakes are not harmful for anyone else. This, however, runs contrary to the way in which suicide affects those emotionally intimate with the suicidal individual. What sort of sensible case can one imagine for a suicide in which no such harm occurs as an effect? If we acknowledge the very reasonable notion that harm exists beyond simply physical

harm, then there is no suicide that doesn't cause psychic harm. With children left behind when their parents commit suicide, the spouses, the coworkers, the friends, the neighbors will all feel some sort of mental or emotional pain at the thought that someone they loved, needed, or even just knew has taken their own life. We can even imagine the hermit with no social or familial ties causes harm to those in society whose empathy is so great that they feel pain simply for the fact that anyone in the world is performing such an act.

One might try to counter this by saying that the amount of harm the act of suicide causes is superseded by the suicidal person's liberty or noninterference rights, at which point the libertarian argument begins to develop the same problem that the utilitarian argument possesses. It seems implausible that physical harm alone can account for the moral permissibility of suicide, and thus the libertarian view likewise begins to look untenable.

While theological philosophy relies on metaphysics that are not universally accepted, utilitarianism on an infeasible means of measuring utility, and libertarianism on individuals' debatable rights, deontology provides a more holistic moral model for judging the permissibility of suicide based on one's moral duties.

Section II: What One's Duties Are, and Which is Greatest

The best account of the permissibility of suicide is one in which the agent is held morally accountable for the actions that they undertake. In the absence of guidance from an omniscient God, one is best guided morally by fulfilling their moral obligations, or duties, to achieve morally good outcomes. Suicide is the sort of action that appears always to cause some harm and always to have the potential to prevent other harm. It is very difficult to conceive of a scenario in which suicide is harm-free. Suppose an elderly woman is suffering from a terminal illness that causes her great

pain and forces her family to care for her in a way that is humiliating to her. One night, she decides to take one pill too many in order to cut her suffering short. Everyone in her family thinks she died of the illness, and not by her own hand, so there is no pain involved in knowing that took her own life. They would be forced to grieve for her either way, so at first it appears that her suicide does not cause any extra harm. In actuality, however, her dishonesty causes everyone who mourns her harm; she cheats them out of an honest understanding of the circumstances of her death. One might try to argue that one isn't harmed if they are unaware of the harm and thus do not feel it. This clearly is not the case, however. If a cashier cheats me by giving short changing me after I pay my bill, I have obviously been harmed regardless of whether or not I notice. In this way we can see that our universal duties must be the anchor to which we stabilize our moral position regarding suicide.

A deontological moral model based on duties, rather than on pleasing God, maximizing utility, or avoiding causing others harm, is therefore the best fit for judging the moral permissibility of suicide. This, of course, means that exploring how suicide interacts with our duties is the necessary next step for determining its permissibility.

The first set of duties whose relationship to suicide we must examine are the duties we have to those with whom we share our lives. These are our duties to our coworkers, our friends, our families, and our spouses; our duties to those we love and our duties to society at large. Of these duties, the most intuitively obvious is the duty to not cause material harm to others. Following this duty would preclude any sort of suicide that had the potential to cause harm to another individual, but suicides in which the self-terminating person doesn't cause any physical harm to another person obviously do not violate this duty. Thus, self-destruction with no other casualties doesn't violate this duty, but a suicide that deals physical damage to others does.

Further, one can see that the next outward duty one has is the

duty to not cause emotional harm to one's loved ones. The notion of giving particular concern to those one is intimate with is hardly controversial. If one's suicide causes shame, sorrow, pity, anguish, or any other sort of emotional pain, then the suicide has interfered with this duty. The possibility of a suicide that doesn't harm one's loved ones in any way will be discussed further in this paper, but for now I will engage with such acts that do bring this sort of harm. While it may seem that this duty alone is enough to declare suicide morally impermissible, we have to remember the point that was made against utilitarianism: that it is difficult to disentangle the emotional harm that suicide causes in the people close to the agent from the harm that is caused by society's preconceived perception of suicide. In the case of the former, one's duty not to cause emotional harm is violated, in the latter it is unclear if this sort of violation occurs. One might try to claim that the harm can be entirely attributed to societal impressions, but it is the action of committing suicide over the years that has created society's impression of the act, and thus suicide still carries the moral weight of breaking one's duty.

It seems clear that suicide could also prevent us from other duties in addition to our duty to do no harm to loved ones. One such additional duty is a personal duty to contribute to society. No person has entirely avoided drinking from the societal cup, even if the amount we indulge varies tremendously. From enjoying a piece of music, to walking down a public sidewalk, or even speaking in a language shared by others, we all enjoy at least some of the cultural bounty that the society we live in offers. In this way there is a duty of reciprocity with society. We are only able to enjoy these societal fruits because of human contributions to society. If one ends their own life via suicide, then one is completely unable to continue contributing to society.

One might argue that when a person is in a suicidal mental state, their anguish could prevent them from contributing to society even if they survive. The response to that must be that as a human

being, a person in such a mental state still possesses agency and the ability to self-determine; if they are human they can still choose to contribute. Did not Jean-Dominique Bauby write *Le Scaphandre et le Papillon* in a state of full body paralysis? Could Kurt Cobain have never written another aurally pleasurable song had he not ended his life? Surely people are still able to contribute to society when placed in positions of desperation, and likewise, many who end their own lives would have been capable of continuing their contributions.

A suicidal individual also faces the most unassailable duties one has: the duties to oneself, the first of which is the duty of self-respect. Obeying one's duty of self-respect means respecting one's own human dignity. What it is in the self that one is respecting when they practice self-respect can best be conceived of as one's human dignity. While there are differing accounts as to just what human dignity is, the most compelling is the Kantian notion that human dignity is derived from a person's ability to choose—their free will. If one ends their life, they are preventing themselves from further self-determination, and are literally denying their own dignity. One may wish to imagine a person in crippling physical agony, unable to properly clean or clothe themselves. In such an instance one may be tempted to believe that to die would be to regain one's dignity that has been lost due to the diminished quality of their life. Surely, though, when measuring human dignity, no amount of personal embarrassment can compare to the loss of dignity of having one's ability to self-determine destroyed.

Most critically, one may wish to claim that there is a deontological duty to avoid harm to oneself, and that suicide might end or prevent harm to oneself. Thus, performing such an act, while still contrary to one's duties to others and contrary to one's duty of self-respect, might cancel them out. The problem with this is that one does not have a duty to prevent harm to oneself. The duty to prevent harm only exists towards others. As the libertarian might argue, one can self-harm, if they so desire, but one may not

act in such a way that harms others. The difference is a matter of consent. This can be easily observed with rituals such as ear piercing. One can consent to having their own ears pierced, an act that causes oneself minor harm, but one cannot ethically pierce other people's ears without their consent. Thus, one's duties to others and of self-respect must be upheld for an act to be morally permissible. We must therefore conclude that suicide is morally impermissible.

Section III: When Suicide Appears to be Morally Permissible

Under our earlier definition of suicide (self-destruction performed intentionally) we can recognize instances of such actions that intuitively might seem morally permissible. Typically, these sorts of actions come in one of two forms: the form of the Martyr and the form of the Burden.

The Martyr is a person who commits suicide with the belief that their death will make the world a better place: the Buddhist monk who sets themselves ablaze to convince people that the Vietnam war is morally repugnant; the kamikaze pilot who crashes their plane to help win a war they believe is righteous; the husband who takes his own life so that his spouse will receive a life insurance payout; the earlier mentioned soldier who leaps atop a grenade to protect her comrades-in-arms. All these cases are of actions that are seemingly performed for causes that the agents believe to be noble. Even if it seems gruesomely utilitarian, intuitively it might seem that a single person dying from a grenade blast is a "better" moral outcome that a group of people dying from a grenade blast, and that, therefore, the soldier's leap is actually morally permissible act.

Consequentialist moral theories maintain that the outcomes of any action carry the action's moral weight. While it may seem appealing to look toward outcomes rather than duties to determine

morality, depending on the former solution seems to bind one with unbreakable causal chains. If we look back at the soldiers and the grenade, we can clearly see how this is the case. Let us say the martyr soldier saves four of her fellows. Let us further say that all four of the saved soldiers go home and mercilessly beat their spouses, and more than that, two of them also mercilessly beat and mentally abuse their children. It would seem that our martyr's actions have actually led to morally impermissible consequences, under a consequentialist model we must hold them morally responsible. This kind of outcome may be unlikely, and it may be more common for seemingly good actions to have good consequences. Yet, no actions are ever performed in a vacuum. Any life saved, any country whose destiny is changed, any war which is ended, is merely one link in a chain of cause and effect stretching infinitely into the future. A consequentialist moral model will fail to account for all of the consequences of a Martyr's suicide.

The other type of suicide that might intuitively seem permissible is that of the Burden. The Burden is a person whose continued existence is more of a drain on those around them than it is a boon. It is possible to see that a Burden may wish for their own death. It may even be the case that the Burden's loved ones wish that they were dead themselves because of the physical and emotional care that they require. One could argue that a Burden's suicide does not violate duties to society and loved ones because of the relief it may provide.

The problem, of course, is that the Burden is still failing to uphold their duty to of reciprocity to society. We must again remember that Jean-Dominique Bauby, with the use of nothing but his eyelids, wrote his autobiography and was still able to contribute to society. One might like to claim that Bauby was an extraordinary man who possessed a will that few others could emulate—a will that allowed him to contribute even in dire circumstances. One must remember, however, the origin of human dignity: the ability to freely choose. It might be blindingly difficult to force

oneself to contribute to society when one is experiencing the type of mental or physical anguish that is often experienced by those who contemplate suicide, but as long as they are human, they still have power to self-determine. Difficult or not, as long as a person can choose, they can choose to find a way to contribute.

Section IV: Corner Cases: When Suicide Still Appears Acceptable

While suicide seems to be almost universally morally impermissible, there do appear to be rare cases where one can perform the act while still behaving morally. The two types of morally permissible suicide are the Ascetic Hermit's Suicide and the Well-Earned Euthanasia.

An ascetic is an individual of faith, who has removed themselves from worldly pleasures, such as food beyond nourishment, alcohol, fine material goods, and sex (Deezia 2017). Examples of the ascetic are St. Francis and his followers, or the Sufis. A hermit is a person who has left society entirely, and who lives on their own in the wilderness, removed from humankind and its artifices. The Ascetic Hermit appears to be an individual who no longer has a duty toward others or of reciprocity to society as they have no close relationships, and they take nothing from society. Because they have eliminated many of their deontological duties, the Ascetic Hermit attempts to live a life free of responsibilities to others. The problem with this is that there is no obvious way of confirming if one has fulfilled their obligations to society or not. How can one's intellectual or spiritual contributions to a society be weighed against the convenience of having a language taught to one from youth? What sort of contribution could one make to a society that would measure up to being given the cultural context that would even lead one to determine that they might be able to fulfill certain duties? Because of the failure to epistemically determine whether one has fulfilled their duties to completion, the

hermit cannot reasonably be thought to be morally free from the duties the average person possesses.

The next sort of apparently permissible suicide is the Well-Earned Euthanasia. A Well-Earned Euthanasia might involve a person who wishes to end their life, and who has gained the approval of their social sphere to do so, having performed a sufficient number of contributing acts to society. In such an instance their duty to those close to them is not left unfulfilled by their death and they have already fulfilled their societal obligations. The question then becomes, "how can one tell if what their contributions to society have equaled or been greater than what they took from society?' The answer, of course, is that there doesn't appear to be a reasonable method for determining when and how such a fulfillment could be made. This doesn't prevent the possibility for the Well-Earned Euthanasia to exist, however. Inability to determine which cases of euthanasia are morally permissible doesn't preclude the possibility that some cases are, in fact, permissible based on the moral actions the suicidal agent has performed throughout their life.

Conclusion

In spite of these corner cases, it seems that we must regard suicide as tragic and morally impermissible in all but the rarest of situations. So long as humans have unfulfilled moral obligations, they have a duty to stay in the game. The acceptance of these duties comes from our intuition, even while many intuitions regarding the morality of suicide lead us astray. We must choose between such intuitions by rigorously examining our duties and confirming which cleave to our intuitions while remaining intellectually consistent. The key is having an internally consistent moral model for applying moral judgements. In that regard, no other system fits better than deontology.

Bibliography

Amundsen, D. (1989) "Suicide and Early Christian Values," in *Suicide and Euthanasia: Historical and Contemporary Themes*, B. Brody (Ed.) (Dordrecht, Netherlands: Kluwer), pp. 77-154

Aquinas, Thomas. (1945, originally written 1273) *Summa Theologica*, in Anton Pegis (ED), *Basic Writings of Saint Thomas Aquinas* (New York: Random House)

Cholbi, M. (2011) *Suicide: The Philosophical Dimensions* (Peterborough, Ontario: Broadview)

Deezia, Burabari S. (2017) "Asceticism: A Match Towards the Absolute," *IAFOR Journal of Ethics, Religion &Philosophy* 3(2), pp. 85-98

Fairbairn, G. (1995) *Contemplating Suicide: The Language and Ethics of Self-Harm* (London: Routledge)

Hardwig, J. (1997) "Is There a Duty to Die?", *Hastings Center Report*, 27, pp. 34-42

Kupfer, Joseph. (1990) "Suicide: Its Nature and Moral Evaluation", *Journal of Value Inquiry*, 24, pp. 67–81

Pabst Battin, M. (1996) *The Death Debate. Ethical Issues in Suicide* (Upper Saddle River, NJ: Prentice-Hall)

Pabst Battin, M. (2015) *The Ethics of Suicide: Historical Sources* (Oxford: Oxford University Press)

Teo, Alan (2012) *Modern-Day Hermits: The Story Hikkomori in Japan and Beyond* (YouTube). (Ann Arbor: University of Michigan Center for Japanese Studies)

The Film-as-Philosophy Debate

Marcel Giwargis

Introduction[1]

Film-as-philosophy is an ongoing debate that revolves around the idea that certain films may do philosophical thinking.[2] In his work *On Film*, Stephen Muhall (2008) uses the first four installments in the *Alien* series to throw his weight behind this position. Thomas Wartenberg, another proponent, defends a "moderate pro-cinematic philosophy" position that presents certain criteria to be met for films to properly be said to be doing philosophy (Wartenberg 2011, p.16). The significance in pursuing this line of thought lies in establishing the plausibility of past films contributing and future films continuing to contribute to the world of philosophy. Critics of the "film as philosophy" hypothesis, hereafter referred to as FAP, make various objections, three of which will be identified and discussed herein (Mulhall 2008, p. 2).

The generality objection states that films are incapable of doing philosophy due to the opposing field of views of the two endeavors: while philosophy concerns itself with generalities and universal truths, films, due to their narrative nature, examine and depict particular instances and not universal truths. The rationality objection states that it would be irrational to use film to do philosophy when texts are established and well-suited for the purpose already. The imposition objection makes the claim that a film does not philosophize on its own. Rather, it is the philosopher viewing the film and thinking about it that imposes a philosophical interpretation on the film. This paper will defend Wartenberg's "moderate pro-cinematic philosophy" position against these objections and propose an additional defense against the imposition objection.

In Section I, I present Wartenberg's theory of "moderate pro-

cinematic philosophy." In Section II, his responses to the generality, rationality, and imposition objections will be considered. In Section III, I will proceed to argue that while Wartenberg's position is quite reasonable, it can be stretched further in order to strengthen FAP against the rationality and imposition objections.

Section I: Wartenberg's Theory of Film as Philosophy

The idea that films can philosophize has been around since at least the 70's. Stanley Cavell's *The World Viewed: Reflections on the Ontology of Film* (1971) and Gilles Deleuze's *Cinema 1: L'image-movement* (1983) and *Cinema 2: L'image-temps* (1985) brought this notion to the forefront. Specifically, Deleuze put forth the idea that philosophy, science, and film are all engaged in "inventing," and all equally creative, while maintaining that the three fields are separate entities (Deleuze 1987, p. 318). For Deleuze, it was important to explore and understand the relationship between philosophy and film.[3]

It was not until the early 2000's however, that this idea received greater attention in philosophical circles, with Stephen Muhall (2008) paving the way with his controversial work *On Film*. In 2007, with *Thinking on Screen: Film as Philosophy*, Wartenberg argues that the position taken by Muhall is an "extreme pro-cinematic philosophy," focusing on Mulhall's claim that films "can be seen to engage in systematic and sophisticated thinking about their themes and about themselves—that films can philosophize [in] just the ways that philosophers do" (Mulhall 2008, p. 4).

Wartenberg takes on a moderate pro-FAP position by stating that media in general are incapable of philosophizing, but that the filmic medium, alongside text and oral discourse, can indeed do philosophy when the application is based "upon specific techniques that filmmakers can employ to do philosophy on film, most centrally the thought experiment" (Wartenberg 2016, pp.

167-168). Wartenberg considers this the "local" strategy—the determination of films doing philosophy warrants a case-by-case evaluation, rather than a blanket statement about films as a whole. The three ways that qualify a particular film as doing philosophy are:

(1) The illustration of a philosophical theory (i.e., the film as a referent to the philosophical theory and treatise);

(2) Reflection on the very nature of cinema itself or self-reflexivity;

(3) The portrayal of a thought experiment (TE) or exhibition of thought experimentation.

(Wartenberg 2006, p. 131)

Upon examination, if a particular film displays one of these three criteria, then it can be said to be doing philosophy.

Section II: Objections to Wartenberg's Cinematic Philosophy

The generality objection, at its core, relies on a fundamental difference between philosophy and film, namely the expanse of focus with which the two fields concern themselves. According to Wartenberg, this objection "emphasizes the narrative character of fiction films in contrast to the universalistic aspirations of philosophy" (Wartenberg 2016, p. 169-170). In "The Philosophical Limits of Film" Bruce Russell (2000), who makes use of this objection, claims that "no one can establish that something holds in all possible worlds by presenting an example or two of a possible world depicted in film" (Russell 2000, p. 166). In other words, films are incapable of doing philosophy because they have a narrow field of view or focus, too particular in their concern, while philosophy grapples with broad universalities.

Though Wartenberg admits that this claim of generality poses a challenge to advocates of FAP—they must demonstrate "how

films can replicate philosophy"—he makes use of the thought experiment in his counter-argument. He states that "philosophy, with its interest in establishing universal truths, employs a similar method to narrative fiction films: namely, *'the thought experiment'*"—i.e., TEs are quite particular and non-universalistic narratives in their nature. He goes on to point out that the history of philosophy starting with Plato makes use of these "imaginary scenarios that play a role in the attempt to ground general claims." Wartenberg uses Bernardo Bertolucci's *The Conformist* as an example in which Plato's "Allegory of the Cave" is adapted cinematically to criticize Italian fascism (Wartenberg 2016, pp. 169-170).[4]

Thus, Wartenberg counters the generality objection by pointing out the fallacy in Russell's presupposition that all philosophical undertakings work towards general truths, as TEs clearly do not. That is not to say, however, that TEs strictly function to establish universal necessary truths, and Wartenberg in *Thinking on Screen* (2007) emphasizes as much. One such function of TEs may be to act as a refutation of a general claim, and Russell concedes that films can in fact do this, but considers this ability an insignificant asset for the FAP proponent (Russell 2000, p. 166).

FAP opponents may go on to question whether TEs are viable tools for doing philosophy to begin with, but must concede that TEs are still widely used in contemporary philosophy. The difference between which are *the best* philosophical tools and which are *accepted* is not one in which FAP proponents need to engage. Since TEs are accepted forms of philosophy, then, by extension, cinematic thought experiments extend the reach of philosophical thought experiments. Nevertheless, Wartenberg's appeal to TEs as a defense against the generality objection, though inventive, leaves the door open to another objection, namely the rationality objection.

The rationality objection against FAP comes from a question posed by Paisley Livingston: "If we in fact believe a more

efficient means to our goal is available, would it not indeed be irrational to pass it by?" (Livingston 2006, p. 17). The implication is that it would be irrational to use films to philosophize when we have tools such as texts that are already perfectly suited for the task (Livingston 2009, p. 56).

Wartenberg lacks a direct counter to this argument, but has a nascent response in appealing to the differing capabilities of cinema and traditional texts. His claim that "film is both a visual and temporally extended art form [that] gives it an immediacy that is greater than other art forms in its presentation of philosophy," in conjunction with his analysis of *The Matrix* in *Thinking on Screen*, conveys the idea that cinema has innate abilities that texts lack and can thus philosophize (Wartenberg 2007, p. 137). Wartenberg considers *The Matrix* a thought experiment that, due to cinematic abilities, goes beyond the evil genius of Descartes's writings by placing the viewer in the same doubtful position as the protagonist, Neo. This gives the viewer a personal experience of doubt throughout the film—something that cinema is capable of doing.

Wartenberg is not arguing that films should replace texts as the mode for doing philosophy. Rather, it is that films can complement texts in doing philosophy and extend the reach of traditional methods of philosophy, as *The Matrix* example illustrates. Recall that Wartenberg adopts a "moderate pro-cinematic philosophy" that bases a film's ability to do philosophy on three criteria. Excluding films that merely illustrate a philosophical talk or conversation—i.e., a philosopher discussing a philosophical topic (an extension of text)—we can say that a film does philosophy when we can point to the existence of a recognized text of philosophy on a particular subject that may act as the referent.[5] Wartenberg says as much when he states that one conception of philosophy is that of "a discipline that addresses a rather limited set of what are often termed 'eternal questions'" (Wartenberg 2007, p. 29). In this way, Wartenberg sees FAP as *complementary* to traditional philosophy.

Since the idea of including TEs on film as a complement to texts in doing philosophy relies on the acceptance of film possessing a unique nature, this does not require the moderate FAP proponent to accept what Livingston terms the "bold thesis" (Livingston 2009). This thesis states that due to the unique abilities of cinema, films can contribute to philosophy in original ways. In order to meet this strong requirement, films need to fulfill the following criteria:

(1) The "means condition", which asserts that film is capable of doing philosophy by way of cinematic means which are exclusive to film; and

(2) The "results condition", which requires the philosophy being conducted by the film to be original and innovative (Livingston 2009, p. 11).

Livingston, I think correctly, rejects the bold thesis. To require film to conduct innovative and original philosophy is something that contemporary conventional philosophy itself may fail to do. But to adopt the moderate FAP position, all that is needed for a film to philosophize is a scene that presents a new TE in cinematic terms, adds new premises to an argument, or simply demonstrates ideas from a different angle with respect to a traditional philosophical issue.

In tackling the means condition, let us agree that Steven Spielberg's *Minority Report* (2002) is a film about the free will debate (or, in a slightly different reading, the effects of free will vs. determinism). Clearly, this is not an original idea, however, the film does present a different perspective on the debate with its use of sound, lighting, acting, and narrative. Note that this does not force the FAP advocate to endorse a particular cinematic theory that touts cinema's unique abilities absent in other art forms. Aaron Smuts, in defending the bold thesis, states that the main concern should be "means which are significantly more cinematic than merely presenting a philosophical lecture" (Smuts 2006, p.

410-411). Thus, it is not crucial for *Minority Report* to have a unique essence that is limited to film. On the other hand, if the argument were to state that *Minority Report* is a TE on the effects of free will vs. determinism, the FAP opponent can, at most, reiterate the rationality objection and question whether it is justified to use cinematic TEs over text TEs.

A way out of this is to advocate McClelland's (2011) "Socratic Model." The model states that "film's inability to express *explicit* reasoning or *general* conclusions actually makes it a suitable medium for prompting an audience into reaching philosophical conclusions for themselves" for film "prompts its audience into greater philosophical understanding precisely by not making explicit philosophical claims about narrative, but rather by inviting us to do some of the work for ourselves" (McClelland 2011, p. 26). While one may object that *Minority Report* is not explicit enough and thus does not philosophize, McClelland's idea leaves the door open to the capacity of the audience for drawing its own general conclusions, in vein of the Socratic method. Not only is this an appropriate response to the rationality objection but, as McClelland points out, it is also an epistemic advantage of screen TEs in that they allow us to "use our own capacity for reason to work out the real significance of the scenario" (McClelland 2019, p. 102).

Let us now turn to the "imposition objection:" the claim that a film does not philosophize on its own. Using *Minority Report* as a TE that illustrates the free will debate, the FAP opponent argues that the meaning is imposed onto the film by the philosopher who is interpreting the content and that this meaning was unintended by Spielberg.

In tackling this objection, Wartenberg argues for a "creator-oriented" approach (Wartenberg 2007, p. 25). This reading "attempts to reconstruct the meaning that the author of a work intended" and considers the social background of the author (Wartenberg 2016, p. 175). This is not to be confused with an

"audience-oriented" interpretation, which lacks any restrictions on the intent on the part of the author (ibid, p. 175). Livingston counters Wartenberg by stating that this kind of approach makes it difficult to discern the author's philosophical knowledge and thus leaves room for tremendous imprecision and ambiguity in regards to their intent. To sidestep imposition, Livingston endorses a "partial intentionalism" which imparts partial relevance to the creator's intention (Livingston 2009, p. 84). Instead of taking the author's social background as a basis for their intent, this type of intentionalism considers "internal" and "external" markers to verify the author's intent (Livingston 2009, p. 108). Upon analysis of a film, this partial intentionalism tries to establish whether the presence of philosophy in the film interpreted by the viewer (internal proof) "meshes" with the philosophical ponderings of the author (external proof), as conveyed through interviews and discussions (Livingston 2009. p. 99).

Wartenberg disagrees with this strong stance on intent and instead advocates for the authors to simply demonstrate cognizance of the "philosophical *problèmatique*." According to Wartenberg, "whether a film illustrates a philosophical theory is a factual question that need not depend on the filmmaker's intentions" (Wartenberg 2016, p. 175). In other words, Livingston's partial intentionalism is excessive in determining qualifications of a film for FAP purposes and the fact that one can discern that the free will debate is present in *Minority Report* is sufficient for meeting the criteria for philosophizing. Wartenberg argues that there are various ways of interpretating and considering philosophical themes in a film in concluding whether that film is doing philosophy (Wartenberg 2009, p. 121).

Section III: Combined Intentionalism

Wartenberg states that terms such as "film as philosophy, film-philosophy, filmosophy, and cinematic philosophy" all fall under

the same umbrella and the only differentiating factor is their level of strength in terms of films philosophical abilities – i.e., weak, moderate, and strong (or bold). Wartenberg's moderate FAP stance rests on the following assertions:

(1) Films whose narratives can be considered TEs can do philosophy;

(2) The entertainment value of a work of art does not preclude it from being able to do philosophy;

(3) A creator-oriented interpretation is sufficient in the determination of a film as FAP – i.e., a strong intentionalist account is not necessary;

(4) Collaborative efforts in filmmaking "do not undermine the possibility of the film doing philosophy" (Wartenberg 2016, p. 178).

As discussed previously, for Wartenberg, the only way of gauging a film's philosophical abilities is through a case-by-case analysis, his "local" strategy (Wartenberg 2006, p.131).

However, his third assertion pertaining to intent may be reinforced with the addition of Livingston's "partial intentionalism" (Livingston 2009, p. 84). Where Wartenberg's interpretation leaves the door wide open for virtually any film and filmmaker to be proclaimed to be doing philosophy, Livingston's form of intentionalism narrows the passage and puts greater burden on the part of the author. This kind of combined intentionalism will create a gate with varying openings (depending on the film/filmmaker) that allows entry of individual films as FAP. It permits a greater flexibility in the determination and qualification of films and thus is not too wide, per Wartenberg, nor too narrow, per Livingston.

This combination may also signal the presence of *filmosophers*—i.e., filmmakers that are doing philosophy through film. Through accepting Wartenberg's four assertions in conjunction with his statement that "when I say that a film philosophizes, it

is really a shorthand expression for stating that the film's makers are the ones who are actually doing the philosophy in/on/through film," the creation of a special category of philosophers may be predicated (Wartenberg 2007, p. 12). Thus, in looking at Spielberg's social background and interviews on *Minority Report* (as well as interviews with Philip K. Dick, the writer on whose work the film is based), and analyzing the film itself using Wartenberg's claims and combined intentionalism, Spielberg may turn out to be a *filmosopher* in his work on *Minority Report*.

Building on this framework, we can say that the creators subscribe to a certain philosophy of living. This code, in turn, has its roots in the world of philosophy—the referent of the filmwork. The exploration of such ideas in a character-specific *story-verse* may be characterized as philosophizing in that particular world. The extent to which the creators may philosophize in their narrative approach will totally depend on their philosophical inclinations. *Filmosophers* may include Catherine Breillat, the Coen Brothers, Michael Haneke, Abbas Kiarostami, Krzystof Kieslowki, Stanley Kubrick, Terrence Malick, and Andrei Tarkovsky, to name some prominent writer/directors whose works have established philosophical leanings.[6] This is not to say that filmmakers are philosophers and that they are conducting traditional philosophy. Rather, they are *filmosophers* in that they are using the cinematic mode to explore philosophy-centered themes.

Conclusion

By injecting the Socratic Method into the cinematic thought experiment and by combining intentionalisms, the replies to the rationality and imposition objections are strengthened, buttressing the stance of the FAP proponent. I have argued that certain films can philosophize in new ways using traditional philosophical referents. Though this may seem as erring on the side of caution, it does leave room for expansion of FAP and the inclusion of more

films. Simultaneously, it acts as a foot in the door of philosophy. As Daniel Frampton suggests, "film offers *another future* for philosophy" and "is simply one separate route for philosophy" (Frampton 2006, p. 183-184). In other words, FAP is worth considering as it offers another way of doing philosophy.

Notes

1. I would like to thank Mark Balaguer, Jay Conway, and Michael Shim for their support and encouragement in the writing of this paper, not to mention their tremendous and invaluable help in philosophical and linguistic terms and getting me to think deeper and with greater precision about the underlying arguments and ideas.
2. It is not intended for the terms of film "thinking" or "doing philosophy" to be taken literally but rather metaphorically. In Section 3, it will be clarified that it is the filmmakers that are doing the thinking using film.
3. Deleuze is adamant in his stance that conflating art/film with philosophy is an error. They are distinct fields that have a give and take between them, an interesting relationship that is yet to be fully explored and elucidated.
4. *The Conformist* is the film that brought philosophy as a field of study into my personal view of interest. Viewing the complex visual motifs, the play of light and shadow, and the allegory within the allegory really struck a chord with me and inspired me to pursue philosophy academically. There is something to be said about the display of philosophy in the cinematic medium: the visual and aural nature creates an accessibility to the concepts that language alone does not.
5. The exclusion of films that are merely the presentation of a philosophical lecture is warranted as this can simply be a replacement for actually reading a paper. The only difference is its being recorded on a moving medium with an element of temporality.
6. Ethan Coen has an undergraduate degree in Philosophy from Princeton University; Malick holds a B.A. in Philosophy from Harvard College and began his Master's degree in Philosophy at Magdalen College, Oxford, only to leave without a degree due to a dispute with his advisor over his thesis on Kierkegaard, Heidegger, and Wittgenstein; Haneke studied Philosophy at the University of Vienna.

Bibliography

Deleuze, Gilles. (1987) "What is the Creative Act," in: D. Lapoujade (Ed), *Two Regimes of Madness, Texts and Interviews* 1975-1995, pp. 317-329 (New York, NY: Semiotext(e))

Frampton, Daniel. (2006) *Filmosophy* (New York, NY: Wallflower Press)

Livingston, Paisley. (2006) "Theses on Cinema as Philosophy," *Journal of Aesthetics and Art Criticism*, 64(1), pp. 1-18

_____ (2009) *Cinema, Philosophy, Bergman: On Film as Philosophy* (Oxford: Oxford University Press)

McClelland, Tom. (2011) "The Philosophy of Film and Film as Philosophy," *Cinema* 2, pp. 11-35

_____ (2019) "Film as Philosophical Thought Experiment: Some Challenges and Opportunities," in: C. Rawls, S.Gouveia & D. Neiva (Eds), *Philosophy and Film: Bridging Divides*, pp. 92-116 (New York, NY: Routledge)

Mulhall, Steven. (2008) *On Film*, 2nd Edition (London: Routledge)

Russell, Bruce. (2000) "The Philosophical Limits of Film," *Film and Philosophy*, Special Issue on Woody Allen, pp. 163-167

Smuts, Aaron. (2006) "Film as Philosophy: In Defense of a Bold Thesis," *Journal of Aesthetics and Art Criticism*, 64(1), pp. 409-420

Wartenberg, Thomas E. (2006) "Film as Argument," *Film Studies*, (8), pp. 126-137

_____ (2007) *Thinking on Screen: Film as Philosophy* (London: Routledge)

_____ (2009) "Response to 'My Critics'," *Projections*, (3)1, pp. 117-125.

_____ (2011) "On the Possibility of Cinematic Philosophy," in: H. Carel & G. Tuck (Eds), *New Takes in Film-Philosophy*, pp. 9-24 (London: Palgrave Macmillan)

_____ (2016) "Film as Philosophy: The Pro Position," in: K. Thomson-Jones (Ed), *Current Controversies in Philosophy of Film*, pp. 165-181 (New York, NY: Routledge)

Can Happy Hard Determinists Stay Happy?

James Savage

Introduction

Determinism is the doctrine that every event is causally determined by (or causally necessitated by) prior events together with the laws of physics. Hard determinism (HD) is the thesis that determinism is true and that our intuitive belief in free will is illusory. A Happy Hard Determinist is someone who accepts HD and retains a positive outlook on life. For proponents of Happy Hard Determinsim, the prospect of lacking free will does not necessarily dampen their ambition or desire to live a good life. For those who may be unfamiliar, the problem stems from three intuitive claims that only appear to be so when not considered in conjunction with each other.

(1) We have free will;

(2) Determinism is true;

(3) Determinism and free will are incompatible with one another.

According to HD, things do not *just happen*. There is a reason why some of the pins fell over after you rolled a bowling ball down the lane. Imagine a four-year old walks up and asks you why the pins fell over. You then give an answer regarding the acceleration of the bowling ball and the contact that was made between the ball and the pins. The child, unsatisfied, continues to ask why. At some point, you become mentally exhausted and so tell the child, "They just do!" You know that your response is not true. The pins do not just fall over. You only say things like this because you are too tired to continue answering the child's ques-

tions or, if you have the patience, it is because you no longer know the answer to their question. According to HD, your response is obviously false because there is always a reason for why something happens in the way it does, necessitated by prior events. Without worrying about some kind of paradoxical problem of a chain of causes and events going on forever, there should always be, in theory, answers to the child's questions.

What does this have to do with free will? Well, if HD is tright and every action is necessitated by prior events, then that also includes human actions, and if all human actions are necessarily going to happen because of past events, then it our actions can't be freely willed. But if free will is illusory, then this also seemingly infringes upon the common meaning of moral responsibility. If our decision to steal our neighbor's car was not free, then how could we be responsible for it? Why should we be sent to jail or prison, be forced to pay fines, or be morally reprimanded for actions we could not ultimately control? If we accept that free will and moral responsibility are incompatible with determinism, then HD becomes more plausible. HD solves the problem of free will and determinism by denying premise 1, that is, by denying we have free will. While there are other possible solutions to the problem of free will and determinism for those who remain unconvinced (e.g., libertarianism and compatibilism), for the purpose of this paper, I will be focusing primarily on HD.

My focus in this paper will center around the moral and practical implications of accepting a hard deterministic outlook on moral responsibility and punishment, specifically when considering what to do with people who commit crimes. Were we to fully accept the HD, then our current methods of punishment and imprisonment are in serious need of alteration. I will draw principally from Saul Smilansky (2001) and Derek Pereboom (2017), who are both sympathetic to the idea that HD is true. They merely disagree on whether or not applying a standard hard deterministic model to society, specifically when regarding punishment, quar-

antine, or rehabilitation of people who commit crimes, is even possible or a good thing to do. Pereboom stands on the side of those who would say that society could survive and, in some ways, thrive without the standard conception of free will. Smilansky, by stark contrast, disagrees and argues that HD put into practice would be self-defeating in nature. He even suggests an alternative he calls "illusionism," which states that disabusing people of their illusory beliefs in free will is both bad for society and morally wrong.

Instead of submitting my own unoriginal and superfluous thoughts on whether or not we actually have free will, I will instead join Smilansky and Pereboom in proposing that we in fact do not have free will. By doing so, we can focus on what I think is the much more interesting question of whether or not we could live without it. Smilansky argues that we cannot because of serious consequences he believes would follow from enforcing HD principles when engaging in the criminal justice system. Building on Pereboom, I will argue that we can in fact live without free will, and that Smilansky's quick dismissal of a practical application of HD principles stems from a hasty pessimism that does not allow him to consider other more positive possibilities. I will dispel Smilansky's charges of "self-defeating in nature" levied against HD by providing examples of how an HD society might implement a criminal justice system. In addition to tempering Smilansky's pessimism on the prospects of such a world, I hope to display a functional system that allows for more tolerance and compassion for our fellow human beings.

In Section I, I lay out the two main arguments given from Smilansky against both Pereboom and an HD model of society in general. In Section II, I define Smilansky's concept of "funishment" and describe how it is the key to the entirety of his argument against HD. In Section III, I bring up an alternative to his "funishment" that I argue Smilansky overlooks, specifically quarantine models. By doing so, I show that Smilansky's argument

operates on the false premise that "funishment" would be required in a society that fully accepts a HD outlook on life.

Section I

In his paper, *Pereboom on Punishment: Funishment, Innocence, Motivation, and Other Difficulties*, Saul Smilansky (2017) argues against Derek Pereboom regarding a practical application of HD in modern societies. It is important to note here that Smilansky supposes a form of HD that takes a strong moral stance on punishment, and that other forms of HD may not suppose anything about what to do with criminals. In this paper I will consider two of what I believe to be his most important arguments against this practical application. The first of which is his "Practical Reductio" (PR) argument, which is so named because he argues a practical application of HD would trigger such horrific events that doing so would become untenable. The second involves his concerns regarding utilitarianism, as a possible way out of the conundrum. These problems with utilitarianism to which Smilansky hints refers to the issue of using people as means to an end. Ultimately, I will do away with Smilansky's PR argument which will remove the need to deal with his concerns regarding the standard problems of utilitarianism.

Smilansky opposes HD in practice specifically because of what he believes will happen to punishment as a result. A result of accepting HD is that since people don't have ultimate control over what they do, they cannot be said to be responsible for those actions. So, in this kind of world, how can we justly blame or imprison anyone for their crimes? Smilansky argues that this will lead to what he calls "funishment," and the practice of it will HD to defeat itself. Here I have reconstructed his PR argument (Smilansky 2017, p. 593-594):

1. Admittedly, social conditions that lead to crime ought to be changed—we do not need to accept the current rate at

which we incarcerate and punish people, and we should make greater efforts towards rehabilitation. However, for pragmatic reasons, hard determinists will still have the need to incarcerate wrongdoers away from lawful society despite these people not having free will over their actions.

2. However, doing so will be a great injustice committed upon these people because of their lack of free will.
3. So, the punishment of these people must look very different from the way it looks in today's society, and will thus take the form of "funishment."
4. Institutions of "funishment" will need to be just as secure as today's jails and prisons but will also need to be as delightful as possible, resemble five-star hotels, and inhabitants of these facilities must also have a considerable amount of freedom of movement within the facility. They must not be disallowed from having friends and family visit and even stay with them for extended periods of time. Additionally, occupants of these places must also be adequately compensated for because, according to HD, no one deserves the indignity of being separated from society. This is all in order to counter the grave injustice of restricting these people's freedom of movement for actions that they were not responsible for.
5. In addition to the rise in cost to maintain facilities of this kind, the occurrence of crime will skyrocket due to these much more pleasant institutions failing to instill any kind of deterrence effect in otherwise law-abiding citizens. Would-be criminals in today's society have to balance the risk of prison or jail before committing a crime. The risk is greatly diminished if their only consequences are the accommodations that "funishment" provides.
6. The consequences of HD being fully implemented would be

so horrible (flood of new crime, higher levels of incarceration, extremely high costs of new "funishment" facilities, and a lack of a sense of justice among the general public) that no one could ever desire such a world, not even hard determinists.

7. Thus, the practice of hard determinism is self-defeating.

Smilansky believes that proponents of HD are committed to what he calls "funishment" for moral reasons. He offers another argument for why hard determinists cannot just avoid his argument and pursue a kind of utilitarian-like consequentialism. Summed up, it runs as follows (ibid, p. 596):

1. HD, as a distinct moral position, cannot argue by utilitarian-like consequentialist standards as it would then betray the moral force of HD.
2. This is because making people suffer guilt or punishment merely for socially useful purposes is morally wrong, even more so in a world that accepts that people do not own their actions.
3. Therefore, HD cannot turn to utilitarian-like consequentialism for assistance in overcoming the reductio, for it would thereby completely betray itself as a distinct ethical position.

While I cannot necessarily argue against the traditional problems of utilitarianism, I do believe that Smilansky errs in the same way, as he supposes the hard determinists would when attempting to escape his reductio. He errs when he supposes "illusionism" is the preferred option. This is the preferred option because he believes it will produce better results for society if people do not lose their faith in free will. But who is to say that this practice does not also have nightmarish potential consequences? In this paper I will avoid the problem of utilitarianism by focusing on defeating PR. If I am successful then I will not need to worry about such utili-

tarian problems.

Smilansky states his overall stance on the issue of applying any kind of HD principles to a modern-day society thusly: "HD hence confronts a vicious philosophical fork: attempting to maintain its integrity leads it to failure in practice and to self-defeat even in its own terms; while attempting to avoid those grim outcomes through embracing a utilitarian-like consequentialism leads it to moral and philosophical self-betrayal" (Smilansky 2017, p. 597). Based on this summation of Smilansky's issues with HD, it is then my job to illustrate how HD might be able to retain its integrity without producing these supposed terrible consequences.

Section II

Pereboom, (2001) argues in *Living Without Free Will* argues that accepting HD needn't produce the results that Smilansky supposes. Unfortunately, Pereboom has not written much regarding Smilansky's more unique arguments regarding "funishment." What he does talk about has to do with quarantine/rehabilitation models to replace traditional punishment or Smilansky's "funishment." My main task in this paper is to dismantle Smilansky's idea of "funishment" with some help from Pereboom.

The key to Smilansky's PR is clearly his notion of "funishment." If one can show that "funishment" is unnecessary or that it can be replaced with a viable alternative, then Smilansky's PR will fall apart. Smilansky's biggest worry when it comes to HD is that of "disappearing deterrence." If modern day prisons transition to facilities of "funishment," people will then be less worried about the consequences of their actions if those consequences are as pleasant as Smilansky's "funishment." What reason would a potential thief have to change their mind before robbing a bank when the consequences of doing so are pleasant? Smilansky argues that this would likely result in people committing crimes who would otherwise not do so with strong deterrents like our

modern day prisons. This only follows, however, if the HD theorist must accept "funishment," but I don't that hard determinists are committed to "funishment." There are other viable alternatives to consider. One in particular that Smilansky dismisses far too quickly, one that both Pereboom and I consider, is the quarantine/rehabilitation model. I will argue that such a model can serve as just as much of a deterrent as modern-day prisons.

Smilansky's definition of "funishment" says it "would resemble punishment in that criminals would be incarcerated apart from lawful society; and institutions of funishment would also need to be as secure as current prisons to prevent criminals from escaping" (Smilansky 2017, p. 593). This all seems reasonable and there are reasons to think the HD theorist is committed to something of the sort. His next point is more contentious, however. "… institutions of funishment would also need to be as delightful as possible. They would need to resemble five-star hotels, where the residents are given every opportunity to enjoy life" (ibid, p. 593). Smilansky is right about the practical implications with regard to deterence to deterrence. If "funishment" actually was the only form of incarceration that was available for a hard determinist to practice without contradicting their beliefs on free will and moral responsibility, then the deterrent effect would be greatly diminished, and possibly diminished to such an extent that the society practicing HD would defeat itself. In Smilansky's own language, the society that adopted HD practices would become so "nightmarish" that no one would ever desire such a life.

But hard determinists are not obligated to transition to institutions of "funishment;" a quarantine/rehabilitation model would be sufficient in deterring would-be criminals, a consideration I turn to now.

Section III

Section 3.1: Quarantine as Punishment

In replacing traditional punishment with a kind of quarantine model, we are essentially shifting the blame onto the disease rather than the person. The analogy here is that people with diseases are no more responsible for their afflictions than criminals are for their crimes. Rather than being detained because they deserve it, under this kind of model, criminals are detained pragmatically in order to protect society. In his *Living Without Free Will*, Pereboom (2001) quotes Ferdinand Schoeman who says, "In order to protect society, we have the right to quarantine people who are carriers of severe communicable diseases, then we also have the right to isolate the criminally dangerous to protect society" (Pereboom 2001, p. 174).

Smilansky does not believe that this will solve the problem of disappearing deterrence. He claims that a "punishment as quarantine" model is "in modern times, quite rare, is typically limited in duration, and is mostly imposed on foreigners" (Smilansky 2017, p. 597). From this he reasons that any intuitions we might have about any such models are not well formed and so can't be relied upon. Smilansky claims, specifically, that quarantine is thought to be only justified in cases of immediate danger. Since the majority of crimes are primarily not life-threatening, then quarantine is not typically justifiable (ibid, p. 597). Examples of non-life-threatening crimes would include such crimes as tax fraud, embezzlement, excessive traffic violations, etc. He claims these do not meet the same kind of criteria analogous with why we quarantine for medical reasons. Smilansky would say that the tax evader poses no immediate threat to the lives of other citizens, whereas someone who is actively contagious with COVID-19 does. So, why ought the tax evader be separated from society at all? Here I think Smilansky misses something. Even though in an HD-practicing society one might accept that the tax evader or the

embezzler is not ultimately responsible for their crimes, our intuition that they be separated from society comes doesn't just reduce to immediate risk-to-life considerations, but from our desire to preserve a society that does not suffer from what Smilansky suggests will happen if we do not detain these non-violent criminals. There's an analogy here between a contagious COVID-19 patient and a tax evader. I would argue that the tax evader poses as much of a threat as the contagious person; without a legitimately viable deterrent system in place, they do in fact pose an immediate threat, even if the threat is not to life itself. The threat of looming chaos in the midst of an HD society that will not quarantine non-violent criminals is just as serious as a contagious person spreading their affliction to other people. In a way, the tax evader similarly spreads their "disease" by walking around law-abiding society with impunity. So, a just HD society has an obligation to quarantine this person from society until one could say that they no longer pose any threat to that society, in the same way a COVID-19 patient is made to quarantine for fourteen days before emerging from their confinement. We do this as a way to protect our society and do so without regarding the offender or contagious person as morally responsible for what has happened to them.

Section 3.2: One-Time Crimers

I believe that Smilansky makes a similar mistake when he brings up the issue of what I will call "one-time crimers." He continues with his criticism of quarantine as a form of punishment by saying that we will have no cause to quarantine or detain "one-time crimers." Smilansky describes "one-time crimers" as people who can be said to no longer pose any threat to society after they have committed their crime. One example he gives tells of "children who tamper with a medical device in order to hasten the death of a very wealthy but obnoxious parent from whom they will inherit" (Smilansky 2017, p. 598). In this hypothetical situation, Smila-

nsky describes the child as having always been well-behaved and only having committed this crime out of extreme financial desperation, desperation that is permanently alleviated once the crime is committed. Smilansky claims that the child will have no need of being quarantined in an HD-practicing society because such quarantine is only warranted when people pose an immediate threat to society. Liberatarians and compatibilists have no problems justifying punishment for the "one-time crimers" while the hard determinist, he claims, will be unable to do so.

As I see it, there are two main issues with Smilansky's problem of "one-time crimers." First, I think Smilansky appears to suggest that "one-time crimers" will simply walk free after being found to no longer be potential threats in an HD society. If this were true, then I would agree that this would inspire countless other "one-time crimers." However, it is likely that a process for finding someone to no longer be a potential threat to society (much like our current systems of finding people innocent or guilty) would be quite a lengthy procedure. In addition to the unpleasantness of a long trial, proving oneself to no longer be a potential threat to society would include a litany of criteria. Even in today's society when someone is convicted of a crime and then serves what is deemed to be an appropriate amount of time in prison, they are often (in serious cases) assigned to a parole officer once released. Parole officers work with people who have served time in prison for serious offenses and have since been released, keeping track of these people to make sure they are satisfying the pre-set conditions of their parole. These conditions are likely unpleasant enough to deter crime, as they continue to limit the person's freedom even once they are released. Ideally, these conditions are ones that, if met for a certain amount of time, open the person up to the possibility of regaining their full independence. I see no reason why this same model could not be applied in an HD society to both "one-time crimers" and more serious offenders when the situation calls for it. Those being paroled are

monitored so that they can prove themselves to no longer pose any threat to society.

Clearly, such procedures would be necessary to ensure that none of these supposed "one-time crimers" developed a taste for whatever criminal activity they performed. While these cases may be rare, it is not unthinkable that the child in Smilansky's example may go on to murder more people. This may occur after activating a certain disposition they might have had towards murder that they had never actualized until they did so out of pure desperation. In which case, this child would be quarantined from society, for a particular amount of time deemed appropriate beforehand, in the same way that normal murderers are for the protection of society. This is all done to protect people from potential threats and also to rehabilitate those for reentry to society after committing a crime. This kind of procedure put in place in an HD society for potential offenders would be uncomfortable enough to deter many, much more so than Smilansky's "funishment."

Have we also solved the problem for those who go through some sort of trial to prove that they truly were merely "one-time crimers?" Why should they have to be detained or quarantined if they are found to no longer pose any threat to society? Smilansky notes that the families of those that are victims of tried and true "one-time crimers" will find no sense of justice if they will just be able to walk free after only a long court trial. He continues to say, "the temptations of hatred and inheritance are there for many people, and once it becomes known that 'one-timers' will be unharmed, one-time crimes will become even more tempting" (ibid, p. 598). This brings us back to the second issue with Smilansky's assessment of "one-time crimers." The fact that a lack of deterrent will encourage a devastating amount of people to commit one-time crimes is the justification we need to say that those who commit one-time crimes can be justifiably and pragmatically quarantined for a particular amount of time. Again, this is because the "one-time crimer"—just like the tax evader—who

walks around law-abiding society with impunity will contaminate other citizens who may then cause other people harm by committing those one-time crimes in the same way a COVID-19 positive person might spread the virus to other people. However, it appears that the problem of "one-time crimers" only becomes a problem when there is no morally justified, HD approved, deterrent system in place. In the following sections, I will offer a crude example of how such a quarantine/rehabilitation system might be structured. I believe this is enough to defeat Smilansky's notion of "funishment" and therefore his PR argument, leaving HD ample justification for detaining and quarantining criminals.

Section 3.3: Rehabilitation

While in the case of "one-time crimers" I think the threat of a lengthy court case to prove that one will no longer commit any further crimes will be a decent deterrent in itself, there is still more to say about handling Smilansky's problem of disappearing deterrence. Recalling my main concern of defeating his idea of "funishment," I argued that were we to have a viable alternative to "funishment," then Smilansky's PR would no longer pose a threat to HD. I would now like to turn attention back to Smilansky's picture of "funishment." In describing "funishment," Smilansky essentially describes a kind of luxurious hotel vacation from the world. It is only by describing institutions of "funishment" as pleasantly as he can that his PR has the power it does.

In addition to leaving out certain aspects that would likely take place regarding the punishment and deterrence of criminals in an HD society (like this explored above), Smilansky also makes no mention of any kind of rehabilitative efforts that would absolutely take place within a "funishment" facility, all of which are likely to deter crime and still be justifiable under HD principles. It is absurd to think that when detaining people, even in institutions of "funishment," that there will be no attempt at

rehabilitating these individuals. Even a cursory internet search yields practices that are done in today's prisons that aim at rehabilitation of inmates, and there is no reason why such practices could not occur in a HD prison/quarantine setting. Services such as education, counseling, therapy, and charitable volunteer work are already put in place to help inmates prepare for their eventual release. While I believe it is fair to say that while these services can be extremely rewarding and therapeutic, they can likely also cause certain unpleasantness that could be enough to deter people from crime. There's an analogy here to going to the doctor or the dentist. While most people don't enjoy going to these kinds of places, the majority of people are glad they went and are better for it because they have improved themselves during the process. An HD quarantine system can easily justify the adoption of these practices in order to reform criminals. I think Smilansky would have to agree that this does not betray HD principles because the criminals are never used as a mere means. Rather, as an attempt to reincorporate them into society, they are given every opportunity to better prepare themselves for release without treating them as being morally responsible. Additionally, this kind of quarantine system is preferable to today's prisons because there is no cause to be unnecessarily cruel to any of the people in them. The purpose of criminals being quarantined is based on the preservation of a good society and their eventual release rather than a retributivist and vindictive sense of justice. It is more reasonable than "funishment" because it can actually deter criminals and aims at allowing them the opportunity to regain the rights they had before committing their crime. It may be true that when a person is quarantined for committing a serious crime they may not choose to participate in programs that might help them. It may also be true that such programs will not work for criminals that may be beyond reincorporating back into society. Here I echo a sentiment from Pereboom when considering criminals who are incapable of reform: "when this is not possible, and a criminal must be confined indefinitely,

his life should not be made unnecessarily unpleasant" (Pereboom 2001, pp. 186).

Section 3.4: My Example

I am sure that at this point there may be those who remain convinced that a quarantine system will fail at the deterrence of potential criminals, and that Smilansky is right that this weakening of deterrence means that HD is inapplicable. Here I explore how the quarantine model I have sketched might deter other potential criminals.

Imagine a scenario in this possible HD world where a man (Glenn) is contemplating stealing someone's car. Despite considering the consequences that will follow if he is caught (a long trial, quarantine, loss of certain privileges because of quarantine both during his sentence and after it is concluded, etc.), he ends up going through with it and attempts to steal a car. In the end, Glenn does get caught and, after a trial, is sentenced to quarantine for a certain amount of time, let's say for six months to a year. While quarantined, Glenn has all of his basic needs met, which is nowhere near as expensive as five-star accommodations, is not treated with unnecessary cruelty, has the option to work where he is paid a fair wage (perhaps out in free society while supervised), and also has the opportunity to engage in rehabilitation programs. When released from quarantine, Glenn will have to report to a parole officer to be subjected to further supervision when back in law-abiding society until he can prove he is no longer a threat to society for a particular amount of time.

All of the unpleasantness that Glenn goes through is in order to either help him in his rehabilitation or deter others from committing similar crimes. It is true that different quarantined individuals, depending on severity of crime and their ability to be around other people as in the case of more violent criminals, will have less access to some of these privileges. We are justified in

restricting these rights in Glenn's case, where we take away some of his rights in order to prevent the chaos that might follow from not punishing him.

Section 3.5: Avoiding Smilansky's Problems

All of the disastrous consequences of Smilansky's argument come from his proposition of "funishment" as the only detainment option available in a just HD society. However, I have given a workable alternative to "funishment" in light of which we should reconsider Smilansky's proposed consequences.

Since I have given a potential quarantine model of punishment that can deter crime, the first two of Smilansky's supposed consequences disappear. The threat of new crime can be tempered with a justified detainment of criminals that is just unpleasant enough to be effective without treating the quarantined person as being morally responsible for their actions. With a successful deterrent for potential crime, Smilansky's worry of higher levels of incarceration also disappears.

Regarding the issue of cost, as I have already shown, and Smilansky has agreed, we can confidently say that five-star accommodations are not necessary. Again, however, the lack of a huge increase in incarcerated people that he supposes suggests that the cost to maintain even a two-star quarantine facility would prove to not be catastrophically expensive. These two-star quarantine facilities might resemble (when appropriate to confine a large number of criminals together, such as non-violent offenders) a kind of military or army barracks, which is something that is already funded in our current society by tax dollars. Ultimately, what is important is that the housing of these quarantined people provides their most basic needs, along with amenities such as the rehabilitation programs I discussed earlier, and spaces to visit family and friends, all of which is to help them reenter society without treating them with unnecessary cruelty. It seems unlikely

that this would bankrupt a society to the point of collapse. Smilansky's claim of unbearable cost of a kind of quarantine model or "funishment" stems from a higher level of incarceration, which I believe I have shown to be a non-issue with the quarantine alternative.

Finally, there's the problem of a lack of a sense of justice among the general public. I assume that what Smilansky means by this is that if "funishment" or the quarantine/rehabilitation model is all that awaits criminals, the general public will not be satisfied with that kind of punishment. This likely is because it is the intuition of most of modern society that these criminals should have a much more unpleasant experience in their confinement because they deserve to suffer harsh consequences for their actions. Here I think Smilansky mistakenly attributes intuitions that might not exist in an HD society. Again, I will borrow from Pereboom, where he quotes Albert Einstein in regards to his non-belief in free will as saying, "it has been a continual consolation in the face of life's hardships, my own and others', and an unfailing well-spring of tolerance" (Pereboom 2001, pp. 211). Pereboom surmises from Einstein's words that this is what a HD outlook on life can offer a person. More specifically, he claims that, "what makes this possible, I believe, is that this view can release us from false beliefs that rationalize expressions of moral anger, and thereby diminish such anger and its harmful expressions" (ibid, p. 211). So, the attitudes towards criminals that Smilansky is worried about might not even exist in an HD world. It is likely that, in an HD-practicing society, the owner of the car that Glenn stole shares the kind of tolerance and sympathy that Einstein describes. It is not to say that reactive attitudes like anger and moral resentment would be completely eradicated, but rather that if and when they do occur, in an HD world, these attitudes would be accompanied by an underlying belief that they are not completely justified.

It seems then that Smilansky's supposed catastrophic problems that follow from HD stem only from his own pessimistic

view of it. He admits this himself while advocating for his "illusionism," stating that "this pessimism thus leads me to conservatism on free will. At the edge, it pushes me towards Illusionism" (Smilansky 2017, p. 592). After more accurately describing what a realistic "funishment," taking the form of a quarantine-rehabilitation model, might look like, and then showing how the consequences from his version of "funishment" would not occur in my version, I believe we can conclude that Smilansky's practical reductio argument against a practical application of HD is only possible when he uses his own cherry-picked version of it. In taking care of his PR, Smilansky would now have to hold back on his concerns regarding utilitarian-like consequentialism, since we have seen an example of punishment and deterrence that does not betray hard deterministic principles.

Conclusion

Now, I would like to point out that I do not have all of the answers of how to build and maintain quarantine facilities of the kind that I have described thus far, or whether or not modern-day people would ever be truly receptive to these ideas that have sprung from a hard deterministic outlook. However, my aim in this paper has been to show that Smilansky's "funishment" is not the best that proponents of HD can provide. Again, I argue that if we can have a viable alternative to "funishment," then Smilansky's entire practical reductio against practicing HD no longer poses any threat. Additionally, having this alternative also protects us from Smilansky's utilitarian concerns due to this alternative quarantine model not threatening the principles of the hypothetical HD society that we have supposed. I find it very plausible that hard determinists can develop a quarantine model, similar yet more fine-tuned than the one I have described, that can deter potential criminals without betraying the idea that these people are not ultimately responsible for their crimes. While radically different from the way things are

done today, I believe the biggest difference in transitioning to a hard deterministic practice would be the replacement of anger and moral resentment felt towards criminals with more compassion and tolerance. I do not think these 'consequences' mark the devastation of any society that may attempt practicing HD. Perhaps Smilansky, who I have just accused of extreme pessimism, might accuse me of wishful thinking. Were that to happen I would cheerfully disagree.

Bibliography

Pereboom, Derk. (2001) *Living Without Free Will*. (Cambridge: Cambridge University Press)

Smilansky, Saul. (2017) "Pereboom on Punishment: Funishment, Innocence, Motivation, and Other Difficulties," *Criminal Law, Philosophy* 11, pp. 591-603

Contributors

Stephen Boynton, B.A. Philosophy, University of California, Berkeley. B.A. Cognitive Science, University of California, Berkeley. Philosophy of Mind, Philosophy of Language, Epistemology, and Metaphilosophy. Stephen will apply to Philosophy Ph.D. programs following completion of his M.A degree.

Danny Castro, B.A. English… M.A. Humanities, California State University, Dominguez Hills. Philosophy of Time. Danny plans to apply to Ph.D. programs and find a girlfriend upon completion of his M.A. degree.

David Dixon, B.A. Philosophy, University of California, Los Angeles. Philosophy of Mind, History of Philosophy, Philosophy of Science, Social and Political Philosophy. David will apply to Philosophy Ph.D. programs upon completion of his M.A. degree.

Marcel Giwargis, B.S. Physiological Sciences, University of California, Los Angeles. M.F.A. Film and Television Production, University of Southern California. Directing, Cinematography, Producing. Marcel will apply to law schools and Ph.D. programs upon completion of his M.A. degree.

John Hurley, B.A. English, University of California, Berkeley. M.A. in History, Harvard University. M.S. in Computer Science, California State University, Los Angeles. Ancient Philosophy, Ethics. John is a Lecturer in Computer Science at California State University, Los Angeles.

Alex Meyering, B.A. English, University of California, Santa Barbara. Spinoza, History of Philosophy, Philosophy of Disability. Alex will begin the Ph.D. program in philosophy at University of California, Santa Cruz in fall 2021 after completing their M.A.

Heather Norwood, B.A. English, University of Kentucky. B.A. Philosophy, University of Kentucky. M.A. English Composition, California State University, San Bernardino. Metaphysics, Philosophy of Language, Logic. Heather will apply to Ph.D. programs in Philosophy upon completion of her M.A. degree.

Sabrina Pirzada, B.A. English, Mount Saint Mary's University. Ancient and Medieval Philosophy, Philosophy of Religion, Philosophy of Science. Sabrina will continue her career at Caltech upon completion of her dual M.A. programs in Economics and Philosophy.

David Randall, A.A. Philosophy, East Los Angeles College. Philosophy of Human Rights, Ethics, Metaphysics. David will apply to law schools upon completion of his B.A. degree.

James Savage, B.A. Philosophy, University of California, Santa Barbara. Ethics, Metaphysics. James will apply for jobs in education upon completion of his M.A. degree.

Master of Arts in Philosophy
California State University, Los Angeles

The Department of Philosophy at California State University, Los Angeles offers a program of study leading to the Master of Arts degree in Philosophy. The program aims at the acquisition of a broad background in philosophy. It is designed for those preparing for further graduate study or community college teaching, and for self-enrichment. Although the department is analytically oriented, it encourages work in other areas, for example Asian philosophy, feminist philosophy, and the interaction between European and Anglo-American thought. The Department includes faculty members with diverse backgrounds and interests actively working in a wide range of philosophical specialties. Classes and seminars are small with a friendly, informal atmosphere that facilitates student-faculty interaction.

The academic programs in philosophy at California State University, Los Angeles are intended to engage students in philosophical inquiry. They aim to acquaint students with noteworthy contributions by philosophers to the tradition; to explore various philosophical issues, problems, and questions; to provide students with principles of inquiry and evaluation relevant to the many areas of human activity, such as science, law, religion, education, government, art, and the humanities; to develop in them skills of analysis, criticism, and synthesis needed for advanced work in various scholarly fields; to encourage the development of skills and attitudes leading to self-reflection and life-long learning.

Philosophy in Practice
Submission Information

Each of the student contributors was specially selected to submit a paper for this issue of *Philosophy in Practice* by one or more faculty members in the Department of Philosophy at California State University, Los Angeles. All writers are currently either students in the master's program of philosophy or undergraduate majors in philosophy. All philosophy students at California State University, Los Angeles are eligible for nomination, and those who were chosen to contribute have demonstrated a superior ability to develop and compose works of advanced philosophical writing.

For more information on *Philosophy in Practice*, please contact:
mshim@calstatela.edu